IMPROVING NEWTON'S METHOD PERFORMANCE BY PARAMETRIZATION: THE CASE OF RICHARDS EQUATION

KONSTANTIN BRENNER AND CLÉMENT CANCÈS

ABSTRACT. The nonlinear systems obtained by discretizing degenerate parabolic equations may be hard to solve, especially with Newton's method. In this paper, we apply to Richards equation a strategy that consists in defining a new primary unknown for the continuous equation in order to stabilize Newton's method by parametrizing the graph linking the pressure and the saturation. The resulting form of Richards equation is then discretized thanks to a monotone Finite Volume scheme. We prove the well-posedness of the numerical scheme. Then we show under appropriate non-degeneracy conditions on the parametrization that Newtons method converges locally and quadratically. Finally, we provide numerical evidences of the efficiency of our approach.

Keywords. Richards equation, Finite Volumes, Newton's method, parametrization

AMS subjects classification. 65M22, 65M08, 76S05

1. INTRODUCTION

1.1. **Motivations and presentation of the Richards equation.** Solving numerically some nonlinear partial differential equations, for example by using finite elements or finite volumes, often amounts to the resolution of some nonlinear system of equations of the form:

$$\text{Find } \boldsymbol{u} \in \mathbb{R}^N \text{ such that } \quad \mathcal{R}(\boldsymbol{u}) = \mathbf{0}_{\mathbb{R}^N}, \tag{1}$$

where $N \in \mathbb{N}^*$ is the number of degrees of freedom and can be large. One of the most popular method for solving the systems of the form (1) is the celebrated Newton-Raphson method. If this iterative procedure converges, then its limit is necessarily a solution to (1). However, making the Newton method converge is sometimes difficult and might require a great expertise. Nonlinear preconditioning technics have been recently developed in improve the performance of the Newton's method, see for instance [18, 7].

Complex multiphase or unsaturated porous media flows are often modeled thanks to degenerate parabolic problems. We refer to [3] for an extensive discussion about models of porous media flows. For such degenerate problems, making Newton's method converge is often very difficult. This led to the development of several strategies to optimize the convergence properties, like for instance the so-called continuation-Newton method [45], or trust region based solvers [44]. An alternative approach consist in solving (1) thanks to a robust fixed point procedure with linear convergence speed rather that with the quadratic Newton's method (see for instance [46, 30, 31, 40]). Comparisons between the fixed point and the Newton's

This work was supported by the GeoPor project funded by the French National Research Agency (ANR) with the grant ANR-13-JS01-0007-01 (project GEOPOR) .

strategies are presented for instance in [33, 4] (see also [42]). Combinations of both technics (perform few fixed points iterations before running Newton's algorithm) was for instance performed in [35].

Our strategy consists in reformulating the problem before applying Newton's method. The reformulation consists in changing the primary variable in order to improve the behavior of Newton's method. We apply this strategy to the so called *Richards equation* [43, 3] modeling the unsaturated flow of water within a porous medium. Extension to more complex models of porous media flows will be the purpose of the forthcoming contribution [5].

Denote by Ω some open subset of $\mathbb{R}^d$ ($d \leq 3$) representing the porous medium (in the sequel, Ω will be supposed to be polyhedral for meshing purpose), by $T > 0$ a finite time horizon, and by $Q := \Omega \times (0, T)$ the corresponding space-time cylinder. We are interested in finding a saturation profile $\overline{s} : Q \to [0, 1]$ and a water pressure $\overline{p} : Q \to \mathbb{R}$ such that

$$\partial_t \overline{s} - \boldsymbol{\nabla} \cdot (\lambda(\overline{s}) (\boldsymbol{\nabla} \overline{p} - \mathbf{g})) = 0, \tag{2}$$

where the mobility function $\lambda : [0, 1] \to \mathbb{R}_+$ is a nondecreasing $\mathcal{C}^2$ function that satisfies $\lambda(s \leq 0) = 0$ and $\lambda(s > 0) > 0$, and where $\mathbf{g} \in \mathbb{R}^d$ stands for the gravity vector. In order to ease the reading, we have set the porosity equal to 1 in (2) and neglected the residual saturation. The pressure and the water content are supposed to be linked by some monotone relation

$$\overline{s} = S(\overline{p}), \qquad \text{a.e. in } Q \tag{3}$$

where S is a non-decreasing function from $\mathbb{R}$ to $[0, 1]$. In what follows, we assume that $S(p) = 1$ for all $p \geq 0$, that corresponds to assuming that the porous medium is water wet, and that $S \in L^1(\mathbb{R}_-)$, implying in particular that $\lim_{p \to -\infty} S(p) = 0$. As a consequence of the Lipschitz continuity of λ and of the integrability of S on $\mathbb{R}_-$, one has

$$\lambda(S) \in L^1(\mathbb{R}_-). \tag{4}$$

We denote by $p_\star = \sup\{p \mid S(p) = 0\}$, with the convention that $p_\star = -\infty$ if $\{p \in \mathbb{R} \mid S(p) = 0\} = \emptyset$.

Typical behaviors of λ and S are depicted in Figure 1.

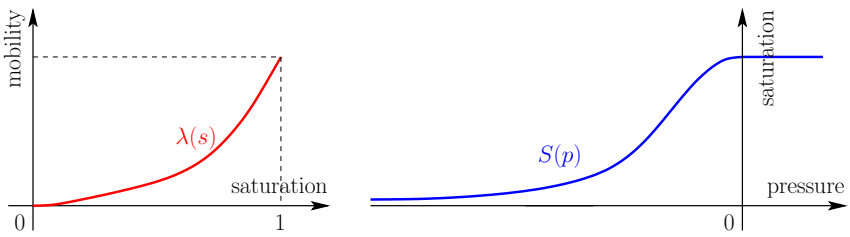


FIGURE 1. The mobility function $\lambda : [0, 1] \to \mathbb{R}_+$ is increasing and satisfies $\lambda(0) = 0$. The saturation function $S : \mathbb{R} \to [0, 1]$ is non-decreasing, constant equal to 1 on $\mathbb{R}_+$ and increasing on $(p_\star, 0)$ for some $p_\star \in [-\infty, 0)$.

Remark 1.1. *In the case where the domain Ω and the finite time T are large, a classical hyperbolic scaling consisting in replacing $(\boldsymbol{x}, t)$ by $(\boldsymbol{x}/\epsilon, t/\epsilon)$ leads to the problem*

$$\partial_t S^\epsilon(\overline{p}) - \boldsymbol{\nabla} \cdot (\lambda(S^\epsilon(\overline{p}))\,(\boldsymbol{\nabla}\overline{p} - \mathbf{g})) = 0,$$

where the function S^ϵ is deduced from S by

$$S^\epsilon(p) = S(p/\epsilon), \qquad \forall p \in \mathbb{R}.$$

Letting ϵ tend to 0 leads the maximal monotone capillary pressure graph

$$S^0(p) = sign_+(p) = \begin{cases} 0 & \text{if } p < 0, \\ [0,1] & \text{if } p = 0, \\ 1 & \text{if } p > 0. \end{cases}$$

The Richards equation then degenerates into a hyperbolic-elliptic problem. Our purpose can be extended to the degenerate case even though the graph S^0 lack regularity thanks to the so-called semi-smooth Newton method (cf. [41]). Moreover, because of the hyperbolic degeneracy, additional entropy criterions à la *Carrillo [13] are required in order to characterize the relevant solution. In order to simplify our purpose as much as possible, we focus on the case $\epsilon > 0$.*

The problem (2)–(3) is complemented by the initial condition

$$\overline{s}_{|_{t=0}} = s_0 \in L^\infty(\Omega; [0,1]), \tag{5}$$

and by Dirichlet boundary conditions on the pressure:

$$\overline{p}_{|_{\partial\Omega\times(0,T)}} = p_D. \tag{6}$$

The regularity requirements on the boundary condition will be specified later on.

At least from a mathematical point of view, it is natural to solve (2) by choosing p or the Kirchhoff transform u (to be defined later on at (8)) as a primary unknown then to deduce $s = S(p) = \widetilde{S}(u)$. However, this approach lacks efficiency when one aims to solve the Richards equation numerically, especially for dry media, i.e, when the saturation s is close to 0. In this latter situation, it turns out that the Newton methods encounters difficulties to converge for solving the nonlinear system obtained thanks to standard implicit numerical methods (say $\mathbb{P}1$-Finite Elements [33], mixed finite elements [4], or Finite Volumes [25, 27]). A better choice as a primary unknown in the dry regions is the saturation s, p or u being computed thanks to the inverse function of S or $\widetilde{S}$ respectively. But choosing the saturation s as the primary variable yields difficulties in the saturated regions, i.e., where $s = 1$. Hence, a classical approach for solving numerically the Richards equation consists in applying the so-called *variable switch*, that consists in changing the primary variable following the physical configuration (see, e.g., [17]).

1.2. **Monotone parametrization of the graph.** The main feature of our contribution consists in parametrizing the graph S in order to stabilize the Newton algorithm without implementing the possibly complex variable switch procedure. This procedure is inspired from the one proposed by J. Carrillo [14] (see also [36] for numerical issues) to deal with hyperbolic scalar conservation laws with discontinuous flux w.r.t. the unknown. Let us introduce two continuously differentiable nondecreasing functions

$$s : (\tau_\star, \infty) \to [0,1] \quad \text{and} \quad p : (\tau_\star, \infty) \to (-\infty, \infty),$$

where $\tau_\star < 0$ may be equal to $-\infty$, such that $p(0) = 0$ and

$$\overline{s} \in S(\overline{p}) \quad \Leftrightarrow \quad \text{there exists } \tau \geq \tau_\star \text{ s.t. } \overline{s} = s(\tau) \text{ and } \overline{p} = p(\tau).$$

This enforces in particular that $\lim_{\tau \to \tau_\star} s(\tau) = 0$ and $\lim_{\tau \to \tau_\star} p(\tau) = p_\star$. In the case where $\tau_\star > -\infty$, the functions s and p are then continuously extended into constants on $(-\infty, \tau_\star)$. It is assumed that the parametrization function s satisfies

$$1 - s \in L^1(\mathbb{R}_+) \text{ and } s \in L^1(\mathbb{R}_-). \tag{7}$$

The Kirchhoff transform $u : [\tau_\star, +\infty) \to \mathbb{R}$ is defined by

$$u(\tau) = \int_0^\tau \lambda(s(a))p'(a)\mathrm{d}a, \qquad \forall \tau \geq \tau_\star. \tag{8}$$

It follows from the integrability property (4) that

$$u_\star := \lim_{\tau \searrow \tau_\star} u(\tau) \quad \text{is finite.} \tag{9}$$

For technical reasons, the function u is artificially extended into a continuous onto function from $\mathbb{R}$ to $\mathbb{R}$ by setting

$$u(\tau) = \tau - \tau_\star + u_\star, \qquad \forall \tau < \tau_\star. \tag{10}$$

However, as it will appear later on (cf. Lemma 2.7), the choice of this extension has no influence on the result.

It is assumed throughout this paper that the parametrization is not degenerated, i.e.,

$$s'(\tau) + p'(\tau) > 0 \quad \text{for a.e. } \tau \geq \tau_\star,$$

or equivalently

$$s'(\tau) + u'(\tau) > 0 \quad \text{for a.e. } \tau \in \mathbb{R}$$

since $\lambda(s(\tau)) > 0$ for all $\tau > \tau_\star$. Since S is an absolutely continuous function, this implies in particular that $p' > 0$ a.e. in $\mathbb{R}_+$.

Such a parametrization of the graph S always exists but is not unique. For instance, on can choose the parametrizations defined by $p(\tau) = \tau$ or $u(\tau) = \tau$. As it will appear in the analysis carried out in the core of the paper, a convenient parametrization should satisfy: there exist $\alpha_\star > 0$ and $\alpha^\star \geq \alpha_\star$ such that

$$\alpha_\star \leq \max(s'(\tau), u'(\tau)) \leq \alpha^\star, \qquad \forall \tau \in \mathbb{R}. \tag{11}$$

Thus, we assume that (11) holds for the analysis. This ensures in particular that the functions s and u are Lipschitz continuous:

$$\|s'\|_\infty \leq \alpha^\star, \qquad \|u'\|_\infty \leq \alpha^\star. \tag{12}$$

For technical reasons that will appear in the analysis, we also assume that there exists $C > 0$ such that

$$\tau \leq C(u(\tau) + 1), \qquad \forall \tau \geq 0. \tag{13}$$

It is also assumed that

$$\liminf_{\tau \searrow \tau_\star} p'(\tau) > 0. \tag{14}$$

This assumption is very naturally satisfied for any non-degenerate parametrization in the sense of (11) of a reasonable function S, but unphysical counterexamples can be designed, enforcing us to set (14) as an assumption.

The function s from $[\tau_\star, 0]$ to $[0,1]$ is nondecreasing and onto. Therefore, one can define the function $s^{-1} : [0,1] \to [\tau_\star, 0]$ by

$$s^{-1}(a) = \min\{x \geq 0 \mid s(x) = a\}, \qquad \forall a \in [0,1]. \tag{15}$$

This allows to define an initial data τ_0 as $\tau_0 = s^{-1}(s_0)$ such that $s(\tau_0) = s_0$.

Choosing τ as the primary variable leads to the following doubly degenerate parabolic equation

$$\partial_t s(\tau) - \boldsymbol{\nabla} \cdot \Big(\lambda(s(\tau)) \big(\boldsymbol{\nabla} p(\tau) - \mathbf{g} \big) \Big) = 0 \quad \text{in } Q.$$

This equation turns to

$$\partial_t s(\tau) + \boldsymbol{\nabla} \cdot \Big(\lambda(s(\tau)) \mathbf{g} - \boldsymbol{\nabla} u(\tau) \Big) = 0 \quad \text{in } Q, \tag{16}$$

at least if $\tau \geq \tau_\star$ (this will be ensured, cf. Theorem 1.3). It is relevant to impose the boundary condition

$$\tau_{|_{\partial\Omega \times (0,T)}} = p^{-1}(p_D) =: \tau_D \geq \tau_\star. \tag{17}$$

as a counterpart of (6). It is finally assumed that τ_D can be extended to the whole $\Omega \times (0,T)$ in a way such that

$$\tau_D \in C^1(\overline{Q}), \qquad \text{with} \quad \tau \geq \tau_\star. \tag{18}$$

The regularity required on τ_D is not optimal and can be relaxed. However, the treatment of the boundary condition is not central in our purpose, hence we stick to (18)

Definition 1.2. *A measurable function $\tau : Q \to \mathbb{R}$ is said to be a weak solution to the problem (16), (5), (17) if $u(\tau) - u(\tau_D) \in L^2\big((0,T); H^1_0(\Omega)\big)$, if $\partial_t s(\tau) \in L^2\big((0,T); H^{-1}(\Omega)\big)$, and if, for all $\varphi \in C^\infty_c(\Omega \times [0,T); \mathbb{R})$, one has*

$$\iint_Q s(\tau) \partial_t \varphi \mathrm{d}\boldsymbol{x}\mathrm{d}t + \int_\Omega s_0 \varphi(\boldsymbol{x}, 0) \mathrm{d}\boldsymbol{x} + \iint_Q \Big(\lambda(s(\tau)) \mathbf{g} - \boldsymbol{\nabla} u(\tau) \Big) \cdot \boldsymbol{\nabla} \varphi \mathrm{d}\boldsymbol{x}\mathrm{d}t = 0.$$

The following statement summarizes known results about the weak solutions.

Theorem 1.3. *There exists a unique weak solution $\tau : \Omega \to \mathbb{R}$ to the problem (16), (5), (17) in the sense of Definition 1.2. Moreover, $\tau \geq \tau_\star$ a.e. in Q, $s(\tau) \in C([0,T]; L^p(\Omega))$ for all $p \in [1, \infty)$, and, given two solutions $\tau, \widehat{\tau}$ corresponding to two initial data s_0 and $\widehat{s}_0$, we have*

$$\int_\Omega \big(s(\tau(\boldsymbol{x}, t)) - s(\widehat{\tau}(\boldsymbol{x}, t)) \big)^\pm \mathrm{d}\boldsymbol{x} \leq \int_\Omega \big(s_0(\boldsymbol{x}) - \widehat{s}_0(\boldsymbol{x}) \big)^\pm \mathrm{d}\boldsymbol{x}, \qquad \forall t \in [0,T]. \tag{19}$$

Existence of weak solutions have been proved by by Alt and Luckhaus in their seminal paper [1]. We refer to [39] (see also [12, 28]) for extended details on the uniqueness proof and on the comparison principle (19). The time continuity of the saturation can be proved as in [8].

1.3. **Outline of the paper.** In §2, we present an implicit monotone Finite Volume scheme [23] designed for approximating the entropy solution τ of (5) and (16)–(17). First, we describe in §2.1 how the domains Ω (and then Q) has to be meshed. In particular, the mesh has to fulfill the so-called *orthogonality condition* so that the diffusion fluxes can be discretized using a simple two-point flux approximation [22].

The Finite Volume scheme is described in §2.2. This scheme yields a nonlinear system of equations

$$\mathcal{F}_n(\boldsymbol{\tau}^n) = \mathbf{0}, \qquad \forall n \in \{1, \dots, N\} \tag{20}$$

to be solved at each time step. The existence and the uniqueness of the solution $\boldsymbol{\tau}^n$ to the nonlinear system (20) is proved at §2.3.

Once we know that the scheme (20) admits one unique solution $\boldsymbol{\tau}^n$, we discuss its effective computation thanks to Newton's method in §3. It is in particular proved in §3.2 that the jacobian matrix is uniformly non-degenerate, so that we can use Newton-Kantorovich theorem to claim the convergence of Newton's method. The quantification of the error linked to the inexact resolution of the nonlinear system is performed in §3.3. Finally, some numerical results are presented in §4 in order to illustrate the efficiency of our approach.

2. The Finite Volume scheme

2.1. **Discretization of Q.**

In this work, we only consider cylindrical discretizations of Q that consist in discretizing space and time separately.

2.1.1. *Admissible mesh of Ω.*

The approximation of the diffusive fluxes we propose relies on the so-called *two-point flux approximation*. This approximation is consistent if the problem is isotropic and if the mesh satisfies the so-called orthogonality condition (see e.g. [22]).

Definition 2.1 (admissible mesh of Ω)**.** *An* admissible mesh $\left(\mathcal{T}, \mathcal{E}, (\boldsymbol{x}_K)_{K\in\mathcal{T}}\right)$ *of Ω is given by a set $\mathcal{T}$ of disjointed open bounded convex subsets of Ω called* control volumes*, a family $\mathcal{E}$ of subsets of $\overline{\Omega}$ called* edges *contained in hyperplanes of $\mathbb{R}^d$ with strictly positive measure, and a family of points $(\boldsymbol{x}_K)_{K\in\mathcal{T}}$ (the so-called* cell centers*). It is assumed that the mesh integrates the whole Ω, i.e., $\bigcup_{K\in\mathcal{T}} \overline{K} = \overline{\Omega}$. The boundary of the control volumes are made of edges, i.e., for all $K \in \mathcal{T}$, there exists a subset $\mathcal{E}_K$ of $\mathcal{E}$ such that $\partial K = \bigcup_{\sigma\in\mathcal{E}_K} \overline{\sigma}$. Furthermore, $\mathcal{E} = \bigcup_{K\in\mathcal{T}} \mathcal{E}_K$. For any $(K, L) \in \mathcal{T}^2$ with $K \neq L$, either the $(d-1)$-dimensional Lebesgue measure of $\overline{K}\cap\overline{L}$ is 0, or $\overline{K}\cap\overline{L} = \overline{\sigma}$ for some $\sigma \in \mathcal{E}$. In the latter case, we write $\sigma = K|L$. We denote by $\mathcal{E}_{\rm int} = \left\{\sigma \in \mathcal{E},\ \exists (K, L) \in \mathcal{T}^2\ \sigma = K|L\right\}$ the set of the internal edges, and by $\mathcal{E}_{\rm ext} = \{\sigma \in \mathcal{E},\ \sigma \subset \partial\Omega\}$, $\mathcal{E}_{K,{\rm ext}} = \mathcal{E}_K \cap \mathcal{E}_{\rm ext}$ of the boundary edges. Finally, the family of points $(\boldsymbol{x}_K)_{K\in\mathcal{T}}$ is such that $\boldsymbol{x}_K \in K$ (for all $K \in \mathcal{T}$) and, if $\sigma = K|L$, it is assumed that the straight line $(\boldsymbol{x}_K, \boldsymbol{x}_L)$ is orthogonal to σ. For all $\sigma \in \mathcal{E}_{\rm ext}$, there exists one unique cell K such that $\sigma \in \mathcal{E}_K$. Then we denote by $\boldsymbol{x}_\sigma$ the projection of $\boldsymbol{x}_K$ over the hyperplane containing σ, and we assume that $\boldsymbol{x}_\sigma$ belongs to σ.*

In what follows, we denote by m_K the d-dimensional Lebesgue measure of the control volume $K \in \mathcal{T}$, and by m_σ the $(d-1)$-Lebesgue measure of the edge $\sigma \in \mathcal{E}$. For all $\sigma \in \mathcal{E}_K$, we denote by $d_{K,\sigma} = d(\boldsymbol{x}_K, \boldsymbol{x}_\sigma)$. Since $\sigma = K|L$ is supposed to be orthogonal to $\boldsymbol{x}_K - \boldsymbol{x}_L$, then $d(\boldsymbol{x}_K, \boldsymbol{x}_L) = d_{K,\sigma} + d_{L,\sigma} =: d_\sigma$. We define the *transmissibilities* $(A_\sigma)_{\sigma\in\mathcal{E}}$ by

$$A_\sigma = \begin{cases} \frac{m_\sigma}{d_\sigma} & \text{if } \sigma = K|L \in \mathcal{E}_{\rm int}, \\ \frac{m_\sigma}{d_{K,\sigma}} & \text{if } \sigma \in \mathcal{E}_K \cap \mathcal{E}_{\rm ext}. \end{cases}$$

The space of the degrees of freedom (including those prescribed by the boundary condition) is

$$\mathbb{X}_{\mathcal{T}} = \left\{ \boldsymbol{v} = (v_K, v_\sigma)_{K\in\mathcal{T},\sigma\in\mathcal{E}_{\rm ext}} \right\} \simeq \mathbb{R}^{\#\mathcal{T}+\#\mathcal{E}_{\rm ext}},$$

while the interior degrees of freedom (for which a nonlinear system has to be solved) are described by the space

$$\mathbb{X}_{\mathcal{T},\rm int} = \left\{ \boldsymbol{v} = (v_K)_{K\in\mathcal{T}} \right\} \simeq \mathbb{R}^{\#\mathcal{T}}.$$

2.1.2. *Time and space-time discretizations.*

Definition 2.2 (Time discretizations of $(0,T)$)**.** *A time discretization of $(0,T)$ is given by an integer value N and a sequence of real values $0 = t^0 < t^1 < \ldots < t^N = T$. For all $n \in \{1,\ldots,N\}$ the time step is defined by $\Delta t^n = t^n - t^{n-1}$.*

Definition 2.3 (Space-time discretizations of Q)**.** *A space-time discretization $\mathcal{D}$ of Q is a family*

$$\mathcal{D} = (\mathcal{T}, \mathcal{E}, (\boldsymbol{x}_K)_{K\in\mathcal{T}}, (t^n)_{n\in\{0,\ldots,N\}}),$$

where $(\mathcal{T}, \mathcal{E}, (\boldsymbol{x}_K)_{K\in\mathcal{T}})$ is an admissible mesh of Ω in the sense of Definition 2.1 and $(N, (t^n)_{n\in\{0,\ldots,N\}})$ is a discretization of $(0,T)$ in the sense of Definition 2.2.

The spaces of the degrees of freedom are defined by

$$\mathbb{X}_{\mathcal{D}} = \left\{ \boldsymbol{v} = (v_K^n, v_\sigma^n)_{K\in\mathcal{T},\sigma\in\mathcal{E}_{\rm ext},1\le n\le N} \right\} \simeq \mathbb{R}^{(\#\mathcal{T}+\#\mathcal{E})\times N}. \tag{21}$$

and

$$\mathbb{X}_{\mathcal{D},\rm int} = \left\{ \boldsymbol{v} = (v_K^n)_{K\in\mathcal{T},1\le n\le N} \right\} \simeq \mathbb{R}^{\#\mathcal{T}\times N}. \tag{22}$$

Let $\boldsymbol{v} = (v_K^n)_{K,n} \in \mathbb{X}_{\mathcal{D}}$, then we denote by $\boldsymbol{v}^n = (v_K^n)_K \in \mathbb{X}_{\mathcal{T}}$ for $n \in \{1,\ldots,N\}$.

2.1.3. *Reconstruction operators.* Following the approach proposed in [20], we introduce reconstruction operators. First, we define the linear operator $\pi_{\mathcal{T}} : \mathbb{X}_{\mathcal{T}} \to L^\infty(\Omega)$ by

$$\pi_{\mathcal{T}}\boldsymbol{v}(\boldsymbol{x}) = v_K \quad \text{if } \boldsymbol{x} \in K, \qquad \forall \boldsymbol{v} = (v_K, v_\sigma)_{K\in\mathcal{T},\sigma\in\mathcal{E}_{\rm ext}}.$$

It is extended into the time-and-space reconstruction linear operator $\pi_{\mathcal{D}} : (\mathbb{X}_{\mathcal{T}})^N \to L^\infty(Q)$ by setting

$$\pi_{\mathcal{D}}\boldsymbol{v}(\boldsymbol{x},t) = v_K^n \quad \text{if } (\boldsymbol{x},t) \in K \times (t^{n-1}, t^n], \qquad \forall \boldsymbol{v} = (v_K^n, v_\sigma^n)_{K\in\mathcal{T},\sigma\in\mathcal{E}_{\rm ext}}^{1\le n\le N} \in (\mathbb{X}_{\mathcal{T}})^N.$$

The study to be performed also requires the introduction of a so-called discrete gradient. We will remain sloppy about the construction of the discrete gradient. We only highlight the properties we will use in the sequel.

Lemma 2.4. *Let $\mathcal{T}$ be an admissible discretization of Ω in the sense of Definition 2.1. There exists a linear operator $\boldsymbol{\nabla}_{\mathcal{T}} : \mathbb{X}_{\mathcal{T}} \to L^\infty(\Omega)^d$ such that for all $\boldsymbol{v} = (v_K, v_\sigma)_{K,\sigma}$ and $\boldsymbol{w} = (w_K, w_\sigma)_{K,\sigma}$ in $\mathbb{X}_{\mathcal{T}}$, one has*

$$\int_\Omega \boldsymbol{\nabla}_{\mathcal{T}}\boldsymbol{v} \cdot \boldsymbol{\nabla}_{\mathcal{T}}\boldsymbol{w} \mathrm{d}\boldsymbol{x} = \sum_{\sigma=K|L\in\mathcal{E}_{\rm int}} A_\sigma (v_K - v_L)(w_K - w_L) + \sum_{K\in\mathcal{T}} \sum_{\sigma\in\mathcal{E}_{K,\rm ext}} A_\sigma (v_K - v_\sigma)(w_K - w_\sigma). \tag{23}$$

Moreover, let $(\mathcal{T}_m)_{m\geq 1}$ *be a sequence of admissible discretizations of* Ω *in the sense of Definition 2.1 such that* $\mathrm{size}(\mathcal{T}_m)$ *tends to* 0 *while* $\mathrm{reg}(\mathcal{T}_m)$ *remains bounded as* m *tends to* ∞*, and let* $(\boldsymbol{v}_m)_{m\geq 1}$ *be a family such that* $\boldsymbol{v}_m \in \mathbb{X}_{\mathcal{T}_m}$ *for all* $m \geq 1$ *and*

$$\|\pi_{\mathcal{T}}\boldsymbol{v}_m\|_{L^2(\Omega)} + \|\boldsymbol{\nabla}_{\mathcal{T}_m}\boldsymbol{v}_m\|_{L^2(\Omega)^d} \leq C, \qquad \forall m \geq 1,$$

then there exists $v \in H^1(\Omega)$ *such that, up to an unlabeled subsequence, one has*

$$\pi_{\mathcal{T}_m}\boldsymbol{v}_m \underset{m\to\infty}{\longrightarrow} v \quad \text{in } L^2(\Omega), \tag{24}$$

and

$$\boldsymbol{\nabla}_{\mathcal{T}_m}\boldsymbol{v}_m \underset{m\to\infty}{\longrightarrow} \boldsymbol{\nabla} v \quad \text{weakly in } L^2(\Omega)^d. \tag{25}$$

Additionally, if $\varphi \in C_c^\infty(\Omega)$ *is discretized into* $\boldsymbol{\varphi}_m = (\varphi_K)_{K\in\mathcal{T}_m} \in \mathbb{X}_{\mathcal{T}_m}$ *by setting*

$$\varphi_K = \frac{1}{m_K}\int_K \varphi(\boldsymbol{x})\mathrm{d}\boldsymbol{x}, \qquad \forall K \in \mathcal{T}_m,$$

then $\boldsymbol{\nabla}_{\mathcal{T}_m}\boldsymbol{\varphi}_m$ *converges strongly in* $L^2(\Omega)^d$ *towards* $\boldsymbol{\nabla}\varphi$ *as* m *tends to* $+\infty$*.*

Note that the discrete gradient reconstruction used in [2, §4.1] (which is the usual one for two-point flux approximations of diffusion operators) does not meet the requirements of Lemma 2.4 since (23) is not fulfilled. However, a reconstruction as prescribed by Lemma 2.4 can be obtained as a particular case of the so-called SUSHI scheme on "super-admissible meshes" (cf. [24, Lemma 2.1]).

Finally, the reconstruction operators $\pi_{\mathcal{T}} : \mathbb{X}_{\mathcal{T}} \to L^\infty(\Omega)$ and $\boldsymbol{\nabla}_{\mathcal{T}} : \mathbb{X}_{\mathcal{T}} \to L^\infty(\Omega)^d$ are extended to the space-times framework into $\pi_{\mathcal{D}} : \mathbb{X}_{\mathcal{D}} \to L^\infty(Q)$ and $\boldsymbol{\nabla}_{\mathcal{D}} : \mathbb{X}_{\mathcal{D}} \to L^\infty(Q)^d$ defined for all $\boldsymbol{v} = (\boldsymbol{v}^n)_{1\leq n\leq N} \in \mathbb{X}_{\mathcal{D}}$ by

$$\pi_{\mathcal{D}}\boldsymbol{v}(\cdot,t) = \pi_{\mathcal{T}}\boldsymbol{v}^n, \qquad \boldsymbol{\nabla}_{\mathcal{D}}\boldsymbol{v}(\cdot,t) = \boldsymbol{\nabla}_{\mathcal{T}}\boldsymbol{v}^n, \qquad \forall t \in (t^{n-1},t^n]. \tag{26}$$

2.2. **The implicit finite volume scheme.** The initial data s_0 is discretized into $\boldsymbol{s}^0 = \left(s_K^0\right)_{K\in\mathcal{T}} \in \mathbb{X}_{\mathcal{T},\mathrm{int}}$ by setting

$$s_K^0 = \frac{1}{m_K}\int_K s_0(\boldsymbol{x})\mathrm{d}\boldsymbol{x}, \qquad \forall K \in \mathcal{T}. \tag{27}$$

Notice that $0 \leq s_K^0 \leq 1$ since $0 \leq s_0 \leq 1$. We define $\boldsymbol{\tau}^0 = \left(\tau_K^0\right)_K \in \mathbb{X}_{\mathcal{T},\mathrm{int}}$ as

$$\boldsymbol{\tau}^0 = s^{-1}(\boldsymbol{s}^0), \tag{28}$$

so that

$$s(\tau_K^0) = s_K^0, \qquad \forall K \in \mathcal{T}.$$

The boundary condition $\tau_D \in C^1(\overline{Q})$ is discretized by setting for all $n \in \{0,\dots,N\}$

$$\tau_{D,\sigma}^n = \tau_D(\boldsymbol{x}_\sigma,t^n), \quad \forall \sigma \in \mathcal{E}_{\mathrm{ext}}, \quad \text{and} \quad \tau_{D,K}^n = \tau_D(\boldsymbol{x}_K,t^n), \quad \forall K \in \mathcal{T}. \tag{29}$$

Then we denote by $\boldsymbol{\tau}_D^n = \left(\tau_{D,K}^n, \tau_{D,\sigma}^n\right)_{K\in\mathcal{T},\sigma\in\mathcal{E}_{\mathrm{ext}}} \in \mathbb{X}_{\mathcal{T}}$, and by $\boldsymbol{\tau}_D = (\boldsymbol{\tau}_D^n)_{1\leq n\leq N} \in \mathbb{X}_{\mathcal{D}}$. It follows from Formula (23) and from the regularity of τ_D that

$$\begin{aligned}\int_\Omega |\boldsymbol{\nabla}_{\mathcal{T}}\boldsymbol{\tau}_D^n|^2\mathrm{d}\boldsymbol{x} &= \sum_{\sigma=K|L\in\mathcal{E}_{\mathrm{int}}} A_\sigma(\tau_{D,K}^n - \tau_{D,L}^n)^2 + \sum_{\sigma\in\mathcal{E}_{\mathrm{ext}}} A_\sigma(\tau_{D,K}^n - \tau_{D,\sigma}^n)^2 \\ &\leq \|\boldsymbol{\nabla}\tau_D\|_\infty^2 \left(\sum_{\sigma=K|L\in\mathcal{E}_{\mathrm{int}}} m_\sigma d_\sigma + \sum_{\sigma\in\mathcal{E}_{\mathrm{ext}}} m_\sigma d_{K,\sigma}\right) = d m_\Omega \|\boldsymbol{\nabla}\tau_D\|_\infty^2.\end{aligned} \tag{30}$$

Let $n \geq 1$. Assume that the state $\boldsymbol{\tau}^{n-1} = \left(\tau_K^{n-1}\right)_K \in \mathbb{X}_{\mathcal{T}}$ is known. The implicit finite volume scheme is obtained by writing the local conservation of the volume of each fluid on the control volumes, i.e.,

$$\frac{s(\tau_K^n) - s(\tau_K^{n-1})}{\Delta t^n} m_K + \sum_{\sigma \in \mathcal{E}_K} F_{K,\sigma}^n = 0, \qquad \forall K \in \mathcal{T}, \tag{31}$$

where $F_{K,\sigma}^n$ denotes the outward w.r.t. K flux across the edge σ at time step t^n. Denote by $\boldsymbol{n}_{K,\sigma}$ the outward w.r.t. K normal to σ, and by $g_{K,\sigma} = \mathbf{g} \cdot \boldsymbol{n}_{K,\sigma}$ for all $\sigma \in \mathcal{E}_K$ and all $K \in \mathcal{T}$. Denote by

$$\tau_{K,\sigma}^n = \begin{cases} \tau_L^n & \text{if } \sigma = K|L \in \mathcal{E}_{\text{int}}, \\ \tau_{D,\sigma}^n & \text{if } \sigma \in \mathcal{E}_{K,\text{ext}}, \end{cases}$$

then the fluxes $F_{K,\sigma}^n$ across $\sigma \in \mathcal{E}_K$ is defined by

$$F_{K,\sigma}^n = m_\sigma \left(\lambda(s(\tau_K^n)) g_{K,\sigma}^+ - \lambda(s(\tau_{K,\sigma}^n)) g_{K,\sigma}^- \right) + A_\sigma \left(u(\tau_K^n) - u(\tau_{K,\sigma}^n) \right). \tag{32}$$

Note in particular that the scheme is locally conservative, i.e.,

$$F_{K,\sigma}^n + F_{L,\sigma}^n = 0, \qquad \forall \sigma = K|L \in \mathcal{E}_{\text{int}}.$$

Combining (31)–(32), the scheme can be rewritten in a condensed form as

$$\mathcal{R}_K \left(\tau_K^n, \tau_K^{n-1}, (\tau_L^n)_{L \neq K}, \left(\tau_{D,\sigma}^n\right)_{\sigma \in \mathcal{E}_{K,\text{ext}}} \right) = 0, \qquad \forall K \in \mathcal{T}, \tag{33}$$

where $\mathcal{R}_K$ is nondecreasing w.r.t. its first argument and nonincreasing w.r.t. the others thanks to the monotonicity of the functions λ, s and u.

It is worth noticing that

$$\sum_{\sigma \in \mathcal{E}_K} m_\sigma g_{K,\sigma} = 0, \qquad \forall K \in \mathcal{T}. \tag{34}$$

Therefore, the convective flux balance can be reformulated, yielding

$$\begin{aligned} \sum_{\sigma \in \mathcal{E}_K} F_{K,\sigma}^n = & \sum_{\sigma \in \mathcal{E}_K L} m_\sigma g_{K,\sigma}^- \left(\lambda(s(\tau_K^n)) - \lambda(s(\tau_{K,\sigma}^n)) \right) \\ & + \sum_{\sigma = K|L} A_\sigma (u(\tau_K^n) - u(\tau_{K,\sigma}^n)), \qquad \forall K \in \mathcal{T}. \end{aligned} \tag{35}$$

2.3. **Existence and uniqueness of the solution to the scheme.** In this section, we analyze the system (33) obtained for a fixed admissible discretization $\mathcal{D}$ of Q. In what follows, we denote by

$$a \top b = \max(a, b) \quad \text{and} \quad a \bot b = \min(a, b), \qquad \forall (a, b) \in \mathbb{R}^2.$$

The following Lemma is a discrete counterpart of the L^1-contraction principle (19) on the exact solution.

Lemma 2.5. *Let $\boldsymbol{\tau}^{n-1}$ and $\widehat{\boldsymbol{\tau}}^{n-1}$ be two elements of $\mathbb{X}_{\mathcal{T},\text{int}}$, and let $\boldsymbol{\tau}^n$ and $\widehat{\boldsymbol{\tau}}^n$ in $\mathbb{X}_{\mathcal{T},\text{int}}$ be two corresponding solutions, then*

$$\int_\Omega |\pi_{\mathcal{T}} s(\boldsymbol{\tau}^n) - \pi_{\mathcal{T}} s(\widehat{\boldsymbol{\tau}}^n)| \, \mathrm{d}\boldsymbol{x} \leq \int_\Omega \left| \pi_{\mathcal{T}} s(\boldsymbol{\tau}^{n-1}) - \pi_{\mathcal{T}} s(\widehat{\boldsymbol{\tau}}^{n-1}) \right| \mathrm{d}\boldsymbol{x}. \tag{36}$$

Proof. It follows from the monotonicity of $\mathcal{R}_K$ that

$$\mathcal{R}_K\left(\tau_K^n, \tau_K^{n-1}\top\widehat{\tau}_K^{n-1}, (\tau_L^n\top\widehat{\tau}_L^n)_{L\neq K}, (\tau_{D,\sigma})_{\sigma\in\mathcal{E}_{K,\mathrm{ext}}}\right) \leq 0,$$
$$\mathcal{R}_K\left(\widehat{\tau}_K^n, \tau_K^{n-1}\top\widehat{\tau}_K^{n-1}, (\tau_L^n\top\widehat{\tau}_L^n)_{L\neq K}, (\tau_{D,\sigma})_{\sigma\in\mathcal{E}_{K,\mathrm{ext}}}\right) \leq 0.$$

Since $\tau_K^n\top\widehat{\tau}_K^n$ is either equal to τ_K^n or to $\widehat{\tau}_K^n$, one has

$$\mathcal{R}_K\left(\tau_K^n\top\widehat{\tau}_K^n, \tau_K^{n-1}\top\widehat{\tau}_K^{n-1}, (\tau_L^n\top\widehat{\tau}_L^n)_{L\neq K}, (\tau_{D,\sigma})_{\sigma\in\mathcal{E}_{K,\mathrm{ext}}}\right) \leq 0. \tag{37}$$

Similar calculations lead to

$$\mathcal{R}_K\left(\tau_K^n\perp\widehat{\tau}_K^n, \tau_K^{n-1}\perp\widehat{\tau}_K^{n-1}, (\tau_L^n\perp\widehat{\tau}_L^n)_{L\neq K}, (\tau_{D,\sigma})_{\sigma\in\mathcal{E}_{K,\mathrm{ext}}}\right) \geq 0. \tag{38}$$

Summing (37) with (38) and over $K\in\mathcal{T}$ yields (36). □

Lemma 2.6. *Given $\boldsymbol{\tau}^{n-1}\in\mathbb{X}_{\mathcal{T}}$, then there exists at most one solution $\boldsymbol{\tau}^n\in\mathbb{X}_{\mathcal{T}}$ to the scheme (31)–(32).*

Proof. As a direct consequence of Lemma 2.5, $s(\tau_K^n) = s(\widehat{\tau}_K^n)$ for all $K\in\mathcal{T}$. Subtracting the system yielding $\widehat{\boldsymbol{\tau}}^n$ to the one corresponding to $\boldsymbol{\tau}^n$ leads to

$$\sum_{\sigma=K|L\in\mathcal{E}_K} A_\sigma\,(w_K^n - w_L^n) + \sum_{\sigma\in\mathcal{E}_{\mathrm{ext}}\cap\mathcal{E}_K} A_\sigma w_K^n = 0, \qquad \forall K\in\mathcal{T},$$

where we have set $w_K^n = u(\tau_K^n) - u(\widehat{\tau}_K^n)$. It follows from classical arguments that $\boldsymbol{w}^n = (w_K^n)_K = \mathbf{0}_{\mathbb{X}_{\mathcal{T}}}$. Bearing the nondegeneracy condition (11) in mind, we get that $\boldsymbol{\tau}^n = \widehat{\boldsymbol{\tau}}^n$. □

The following lemma shows that the solution $\boldsymbol{\tau}^n$ to the scheme is always greater than $\tau^\star$. Therefore, the extension (10) we chose for the function u does not affect the result.

Lemma 2.7. *Let $\boldsymbol{\tau}^n\in\mathbb{X}_{\mathcal{T}}$ be the solution to (31)–(32), then $\tau_K^n\geq\tau_\star$ for all $K\in\mathcal{T}$.*

Proof. There is nothing to prove if $\tau_\star = -\infty$, hence let us assume that $\tau_\star$ is finite. Let K be a cell such that $\tau_K^n\leq\tau_{K,\sigma}^n$ for all $\sigma\in\mathcal{E}_K$, and assume that $\tau_K^n<\tau_\star$. Since $s(\tau_K^n)=0$ and $\lambda(s(\tau_K^n))=0$, one gets that

$$\sum_{\sigma\in\mathcal{E}_K} A_\sigma\left(u(\tau_K^n) - u(\tau_{K,\sigma}^n)\right) \geq 0.$$

The extension (10) of u ensures that $u(\tau_K^n) < u(\tau_\star) \leq u(\tau_{D,\sigma}^n)$. The left-hand side of the above relation is therefore negative, hence a contradiction with the assumption $\tau_K^n<\tau_\star$. □

Let $n\in\{0,\dots,N\}$, then define

$$e_K^n(\tau) = \int_{\tau_{D,K}^n}^{\tau} (a-\tau_{D,K}^n)s'(a)\mathrm{d}a = \int_{\tau_{D,K}^n}^{\tau} \big(s(\tau)-s(a)\big)\mathrm{d}a \geq 0, \quad \forall\tau\in\mathbb{R},$$

and $\mathfrak{E}^n : \mathbb{X}_{\mathcal{T}}\to\mathbb{R}_+$ by

$$\mathfrak{E}^n(\boldsymbol{\tau}) = \sum_{K\in\mathcal{T}} e_K^n(\tau_K)m_K, \qquad \forall\boldsymbol{\tau} = (\tau_K)_{K\in\mathcal{T}}.$$

It is easy to verify (see e.g. [11]) that

$$0\leq\mathfrak{E}^n(\boldsymbol{\tau}) \leq m_\Omega\left(\|1-s\|_{L^1(\mathbb{R}_+)} + \|s\|_{L^1(\mathbb{R}_-)}\right), \qquad \forall\boldsymbol{\tau}\in\mathbb{X}_{\mathcal{T}}. \tag{39}$$

Moreover, it follows from the C^1 regularity of τ_D and from the fact that $0 \leq s \leq 1$ that

$$|\mathfrak{E}^n(\boldsymbol{\tau}) - \mathfrak{E}^{n-1}(\boldsymbol{\tau})| \leq \Delta t^n m_\Omega \|\partial_t \tau_D\|_\infty. \tag{40}$$

We define the Lipschitz continuous function $\xi : \mathbb{R}_+ \to \mathbb{R}$ by

$$\xi(\tau) = \int_0^\tau \sqrt{\lambda(s(a))p'(a)}\mathrm{d}a = \int_0^\tau \sqrt{u'(a)}\mathrm{d}a, \qquad \forall \tau \in \mathbb{R}, \tag{41}$$

then it follows from Cauchy-Schwartz inequality that

$$(a-b)(u(a)-u(b)) \geq (\xi(a)-\xi(b))^2, \qquad \forall (a,b) \in \mathbb{R}^2. \tag{42}$$

Moreover, the Lipschitz continuity of ξ implies that

$$u(\tau) \leq \|\xi'\|_\infty \xi(\tau), \qquad \forall \tau \in \mathbb{R}_+,$$

hence it follows from Assumption (13) that

$$\tau \leq C\left(\xi(\tau)+1\right), \qquad \forall \tau \in \mathbb{R}_+, \tag{43}$$

then, in particular, one has

$$\lim_{\tau \to \infty} \xi(\tau) = +\infty. \tag{44}$$

Lemma 2.8. *Let $\boldsymbol{\tau}^n$ be a solution to the scheme (31)–(32). Then there exists C_1 depending only on Ω and τ_D such that the following estimate holds:*

$$\mathfrak{E}^n(\boldsymbol{\tau}^n) + \Delta t^n \left(C_1 + \frac{1}{2}\int_\Omega |\boldsymbol{\nabla}_{\mathcal{T}} \xi(\boldsymbol{\tau})|^2 \mathrm{d}\boldsymbol{x} \right) \leq \mathfrak{E}^{n-1}(\boldsymbol{\tau}^{n-1}). \tag{45}$$

Proof. We multiply the equation (31) by $\Delta t^n(\tau_K^n - \tau_{D,K}^n)$ and sum over $K \in \mathcal{T}$. Using (35), this provides

$$T_1 + \Delta t^n \big(T_2 + T_3\big) = 0, \tag{46}$$

where

$$\begin{aligned}
T_1 &= \sum_{K \in \mathcal{T}} \left(s(\tau_K^n) - s(\tau_K^{n-1})\right)(\tau_K^n - \tau_{D,K}^n) m_K, \\
T_2 &= \sum_{K \in \mathcal{T}} \sum_{\sigma \in \mathcal{E}_K} m_\sigma g_{K,\sigma}^- \left(\lambda(s(\tau_K^n)) - \lambda(s(\tau_{K,\sigma}^n))\right)(\tau_K^n - \tau_{D,K}^n), \\
T_3 &= \int_\Omega \boldsymbol{\nabla}_{\mathcal{T}} u(\boldsymbol{\tau}^n) \cdot \boldsymbol{\nabla}_{\mathcal{T}}(\boldsymbol{\tau}^n - \boldsymbol{\tau}_D^n) \mathrm{d}\boldsymbol{x}.
\end{aligned}$$

It follows from the convexity of $e \circ s^{-1}$ (see e.g. [9, Proposition 3.7]) that

$$T_1 \geq \mathfrak{E}^n(\boldsymbol{\tau}^n) - \mathfrak{E}^n(\boldsymbol{\tau}^{n-1}).$$

Then thanks to (40), one gets that

$$T_1 \geq \mathfrak{E}^n(\boldsymbol{\tau}^n) - \mathfrak{E}^{n-1}(\boldsymbol{\tau}^{n-1}) - \Delta t^n m_\Omega \|\partial_t \tau_D\|_\infty. \tag{47}$$

The term T_2 can be estimated following the path of [21]. Denote by $\Phi : \mathbb{R} \to \mathbb{R}_+$ the function defined by

$$\Phi(\tau) = \int_0^\tau a s'(a) \lambda'(s(a)) \mathrm{d}a, \qquad \forall \tau \in \mathbb{R}.$$

Since $s(0) = 1$, one has $s'(a) = 0$ for all $a \geq 0$, and thus

$$\Phi(\tau) = 0 \quad \text{if} \quad \tau > 0, \quad \text{and} \quad \Phi'(\tau) \leq 0, \quad \forall \tau \leq 0.$$

Moreover, since $s \in L^1(\mathbb{R}_-)$, one has

$$|\tau|\lambda(s(\tau)) \leq \|\lambda'\|_\infty |\tau| s(\tau) \underset{\tau \searrow \tau_\star}{\longrightarrow} 0.$$

Therefore, for all $\tau \leq 0$, one has

$$\Phi(\tau) = \tau\lambda(s(\tau)) + \int_\tau^0 \lambda(s(a))\mathrm{d}a \leq \int_\tau^0 \lambda(s(a))\mathrm{d}a + C.$$

Thanks to (9) and to Assumption (14), one has $\int_\tau^0 \lambda(s(a))\mathrm{d}a \leq C$, hence Φ is bounded. Simple calculations show that for all $(a,b) \in \mathbb{R}^2$, one has

$$b\left(\lambda(s(b)) - \lambda(s(a))\right) = \Phi(b) - \Phi(a) + \int_a^b \big(\lambda(s(r)) - \lambda(s(a))\big)\mathrm{d}r \geq \Phi(b) - \Phi(a).$$

Rewriting

$$T_2 = T_{21} + T_{22} \tag{48}$$

with

$$T_{21} = \sum_{K\in\mathcal{T}} \sum_{\sigma\in\mathcal{E}_K} m_\sigma g_{K,\sigma}^- \left(\lambda(s(\tau_K^n)) - \lambda(s(\tau_{K,\sigma}^n))\right) \tau_K^n,$$
$$T_{22} = -\sum_{K\in\mathcal{T}} \sum_{\sigma\in\mathcal{E}_K} m_\sigma g_{K,\sigma}^- \left(\lambda(s(\tau_K^n)) - \lambda(s(\tau_{K,\sigma}^n))\right) \tau_{D,K}^n,$$

we get that

$$\begin{aligned} T_{21} &\geq \sum_{K\in\mathcal{T}} \sum_{\sigma\in\mathcal{E}_K} m_\sigma g_{K,\sigma}^- \left(\Phi(s(\tau_K^n)) - \Phi(s(\tau_{D,K}^n))\right) \\ &\geq \sum_{K\in\mathcal{T}} \sum_{\sigma\in\mathcal{E}_{K,\mathrm{ext}}} m_\sigma \left(g_{K,\sigma}^+ \Phi(s(\tau_K^n)) - g_{K,\sigma}^- \Phi(s(\tau_{D,K}^n))\right). \end{aligned}$$

Using the boundedness of Φ, we get that

$$T_{21} \geq -m_{\partial\Omega}|\mathbf{g}|\|\Phi\|_\infty. \tag{49}$$

On the other hand, a classical reorganization of the term T_{22} provides

$$\begin{aligned} T_{22} = &\sum_{\sigma=K|L\in\mathcal{E}_{\mathrm{int}}} m_\sigma \left(g_{K,\sigma}^+ \lambda(s(\tau_K^n)) - g_{L,\sigma}^+ \lambda(s(\tau_L^n))\right)(\tau_{D,K}^n - \tau_{D,L}^n) \\ &- \sum_{K\in\mathcal{T}} \sum_{\sigma\in\mathcal{E}_{K,\mathrm{ext}}} m_\sigma \left(g_{K,\sigma}^+ \lambda(s(\tau_K^n)) - g_{K,\sigma}^- \lambda(s(\tau_{D,\sigma}^n))\right) \tau_{D,K}^n. \end{aligned}$$

Therefore, it follows from the regularity of τ_D and from the boundedness of λ that

$$T_{22} \geq -|\mathbf{g}|\|\lambda\|_\infty \left(dm_\Omega \|\boldsymbol{\nabla}\tau_D\|_\infty - m_{\partial\Omega}\|\tau_D\|_\infty\right). \tag{50}$$

Finally, it results from (42), from the relation $u' = (\xi')^2$, and from (30) that

$$\begin{aligned} T_3 &\geq \|\boldsymbol{\nabla}_\mathcal{T}\xi(\boldsymbol{\tau}^n)\|_{L^2(\Omega)^d}^2 - \|\boldsymbol{\nabla}_\mathcal{T} u(\boldsymbol{\tau}^n)\|_{L^2(\Omega)^d} \|\boldsymbol{\nabla}_\mathcal{T}\boldsymbol{\tau}_D^n\|_{L^2(\Omega)^d} \\ &\geq \|\boldsymbol{\nabla}_\mathcal{T}\xi(\boldsymbol{\tau}^n)\|_{L^2(\Omega)^d}^2 - \sqrt{\|u'\|_\infty} \|\boldsymbol{\nabla}_\mathcal{T}\xi(\boldsymbol{\tau}^n)\|_{L^2(\Omega)^d} \|\boldsymbol{\nabla}_\mathcal{T}\boldsymbol{\tau}_D^n\|_{L^2(\Omega)^d}. \end{aligned}$$

Therefore, it follows from (30) that

$$T_3 \geq \frac{1}{2}\int_\Omega |\boldsymbol{\nabla}_\mathcal{T}\xi(\boldsymbol{\tau}^n)|^2 \mathrm{d}\boldsymbol{x} - \frac{\|u'\|_\infty dm_\Omega \|\boldsymbol{\nabla}\tau_D\|_\infty}{2}. \tag{51}$$

Putting (47)–(51) in (46) ends the proof of Lemma 2.8. □

Proposition 2.9. *Let $\boldsymbol{\tau}^{n-1} \in \mathbb{X}_{\mathcal{T},\mathrm{int}}$, there exists a unique solution $\boldsymbol{\tau}^n \in \mathbb{X}_{\mathcal{T},\mathrm{int}}$ to the scheme (31)–(32). Moreover, it satisfies $\tau_K^n \geq \tau_\star$ for all $K \in \mathcal{T}$.*

Proof. The uniqueness of the solution was proven at Lemma 2.6 while the fact that $\boldsymbol{\tau}^n \geq \tau_\star$ was the purpose of Lemma 2.7. Therefore, it only remains to show the existence of a solution. It follows from Estimate (45) that

$$\|\boldsymbol{\nabla}_{\mathcal{T}}\left(\xi(\boldsymbol{\tau}^n) - \xi(\boldsymbol{\tau}_D^n)\right)\|_{L^2(\Omega)^d}^2 \leq 2\frac{\mathfrak{E}^{n-1}(\boldsymbol{\tau}^{n-1})}{\Delta t^n} + 2C_1 + 4\|u'\|_\infty \|\boldsymbol{\nabla}\tau_D\|_\infty^2.$$

Since ξ is coercive (44), there exists C depending on the data (among which $\boldsymbol{\tau}^{n-1}$, τ_D, the mesh $\mathcal{T}$ and the time step Δt^n) such that

$$|\tau_K^n| \leq C, \qquad \forall K \in \mathcal{T}.$$

This estimate allows us to make use of a topological degree argument (see [34, 16, 21]) to prove the existence of one solution to the scheme. □

The convergence of the scheme can be proved following the path proposed in [25]. Enhanced convergence properties can be obtained thanks to the recent contribution [19]. But this is not the goal of this paper. We are mainly interested in the practical computation of the approximate solution at fixed discretization parameters. In particular, we focus on the behavior the Newton's method.

3. About the Newton method

The numerical scheme (31)–(32) amounts for all $n \in \{1, \ldots, N\}$ to the nonlinear system

$$\mathcal{F}_n(\boldsymbol{\tau}^n) = \left(f_K^n(\boldsymbol{\tau}^n)\right)_{K\in\mathcal{T}} = \mathbf{0}, \qquad \text{with} \quad \mathcal{F}_n \in \mathcal{C}^2\left(\mathbb{R}^{\#\mathcal{T}}; \mathbb{R}^{\#\mathcal{T}}\right), \tag{52}$$

where

$$f_K(\boldsymbol{\tau}) = (s(\tau_K) - s_K^{n-1}) + \frac{\Delta t^n}{m_K}\sum_{\sigma\in\mathcal{E}_K}\left(m_\sigma g_{K,\sigma}^-(\lambda(s(\tau_K^n)) - \lambda(s(\tau_{K,\sigma}^n))) + A_\sigma\left(u(\tau_K^n) - u(\tau_{K,\sigma}^n)\right)\right).$$

Assume that the Jacobian matrix $\mathbb{J}_{\mathcal{F}_n}(\boldsymbol{\tau}^n)$ of $\mathcal{F}_n$ at $\boldsymbol{\tau}^n$ is not singular (this will be shown for nondegenerate parametrizations (11), cf. Proposition 3.7). Approximating the solutions to the system (20) with the Newton method consists in the construction of a sequence $\left(\boldsymbol{\tau}^{n,k}\right)_{k\geq 0}$ defined by

$$\begin{cases} \boldsymbol{\tau}^{n,0} = \boldsymbol{\tau}^{n-1}, \\ \boldsymbol{\tau}^{n,k+1} = \boldsymbol{\tau}^{n,k} - \left[\mathbb{J}_{\mathcal{F}_n}(\boldsymbol{\tau}^{n,k})\right]^{-1}\mathcal{F}_n(\boldsymbol{\tau}^{n,k}). \end{cases} \tag{53}$$

If the method converges, the solution $\boldsymbol{\tau}^n$ is then defined as

$$\boldsymbol{\tau}^n = \lim_{k\to\infty} \boldsymbol{\tau}^{n,k}. \tag{54}$$

Moreover, the convergence speed is asymptotically quadratic if $\mathcal{F}_n \in \mathcal{C}^2$, i.e.,

$$\|\boldsymbol{\tau}^{n,k} - \boldsymbol{\tau}^n\| \leq C\|\boldsymbol{\tau}^{n,k-1} - \boldsymbol{\tau}^n\|^2, \qquad \forall k \geq k_\star \text{ large enough} \tag{55}$$

where, denoting by $\mathcal{V}(\boldsymbol{\tau}^n)$ a neighborhood of $\boldsymbol{\tau}^n$ in $\mathbb{X}_{\mathcal{T}}$, the quantity C (as well as $k_\star$) depends on

$$\sup_{\boldsymbol{\tau}\in\mathcal{V}(\boldsymbol{\tau}^n)}\left\|\left[\mathbb{J}_{\mathcal{F}_n}(\boldsymbol{\tau})\right]^{-1}\right\| \qquad \text{and} \qquad \sup_{\boldsymbol{\tau}\in\mathcal{V}(\boldsymbol{\tau}^n)}\left\|D^2\mathcal{F}_n(\boldsymbol{\tau})\right\|. \tag{56}$$

Since $\boldsymbol{\tau}^n$ is unknown, a sufficient condition to ensure that (55) holds for some $C > 0$ is the following uniform non-degeneracy condition

$$\sup_{\boldsymbol{\tau} \in \mathbb{R}^{\#\mathcal{T}}} \left\| [\mathbb{J}_{\mathcal{F}_n}(\boldsymbol{\tau})]^{-1} \right\| < \infty. \tag{57}$$

We cite here a simplified version of the so-called Newton-Kantorovich theorem (see, e.g., [32, 37, 38]). We refer to [29] for a quantitative version of the theorem, and to [41] to a non-smooth version.

Theorem 3.1 (Newton-Kantorovich theorem)**.** *Assume that there exists two positive quantities C_2 and C_3 such that*

$$\sup_{\boldsymbol{\tau} \in \mathbb{X}_{\mathcal{T}}} \|\mathbb{J}_{\mathcal{F}_n}(\boldsymbol{\tau})\| \leq C_2 \quad \text{and} \quad \sup_{\boldsymbol{\tau} \in \mathbb{X}_{\mathcal{T}}} \left\| [\mathbb{J}_{\mathcal{F}_n}(\boldsymbol{\tau})]^{-1} \right\| \leq C_3, \tag{58}$$

then there exists $\rho > 0$ such that

$$\|\boldsymbol{\tau}^{n,0} - \boldsymbol{\tau}^n\| \leq \rho \quad \Longrightarrow \quad \boldsymbol{\tau}^{n,k} \underset{k \to \infty}{\longrightarrow} \boldsymbol{\tau}^n.$$

Our strategy in the sequel is to prove that a non-degenerate parametrization (in the sense of (11)) yields estimates (58) in the subordinate matrix 1-norm.

Remark 3.2. *The assumption $\mathcal{F} \in \mathcal{C}^1$ can be relaxed. More precisely, Newton method can be extended to the case where $\mathcal{F}$ is merely semi-smooth [15] following the way proposed by Qi and Sun in [41]. Quadratic convergence is preserved in this nonsmooth case provided $\mathcal{F}_n$ is semi-smooth of order 1, cf.* [41, Theorem 3.2]*. Our study can be extended to this more general case, but, for the sake of simplicity, we have chosen to reduce our presentation to the classical smooth framework.*

3.1. **Some technical lemmas related to M-matrices.** Because of the elliptic degeneracy of the problem when $\tau \geq 0$, we cannot apply the the results of [26] to get estimates on the Jacobian matrix $\mathbb{J}_{\mathcal{F}_n}$. We need to introduce some technical material to circumvent this difficulty. Let us first define that notion of δ-transmissive path (see also [9, 10])

Definition 3.3 (δ-transmissive path)**.** *Let $\delta > 0$, let $\mathbb{A} = (a_{ij})_{1 \leq i,j \leq N} \in \mathcal{M}_N(\mathbb{R})$ and let $(i, j) \in \{1, \dots, N\}$.*

(1) A row-wise δ-transmissive path $\mathcal{P}(i, j)$ of length L from i to j associated to the matrix $\mathbb{A}$ consists in a list $\{k_0, \dots, k_L\}$ with
  (i) $k_0 = i$, $k_L = j$, and $k_p \neq k_q$ if $p \neq q$;
  (ii) for all $p \in \{0, \dots, L-1\}$, one has $a_{k_p k_{p+1}} < -\delta$.

(2) The list $\{k_0, \dots, k_L\}$ is a column-wise δ-transmissive path associated to the matrix $\mathbb{A}$ if it is a row-wise δ-transmissive path associated to $\mathbb{A}^T$.

Definition 3.4 ((δ, Δ)-M-matrices)**.** *Let $\delta, \Delta > 0$ be such that $\Delta > \delta$.*

(1) A matrix $\mathbb{A} = (a_{ij})_{1 \leq i,j \leq N} \in \mathcal{M}_N(\mathbb{R})$ is said to be a row-wise (δ, Δ)-M-matrix if
  (i) for all $i \in \{1, \dots, N\}$, one has $\delta \leq a_{ii} \leq \Delta$, $a_{ij} \leq 0$ for all $j \neq i$ and $\sum_{j=1}^N a_{ij} \geq 0$;
  (ii) the set $\mathcal{I}_\delta(\mathbb{A}) = \left\{ i \in \{1, \dots, N\} \;\middle|\; \sum_{j=1}^N a_{ij} \geq \delta \right\}$ is not empty, and for all $i \in \mathcal{I}_\delta(\mathbb{A})^c = \{1, \dots, N\} \setminus \mathcal{I}_\delta(\mathbb{A})$, there exists a δ-transmissive path $\mathcal{P}(i, j)$ with $j \in \mathcal{I}_\delta(\mathbb{A})$.

(2) A matrix $\mathbb{A} = (a_{ij})_{1\le i,j\le N} \in \mathcal{M}_N(\mathbb{R})$ *is said to be a column-wise* (δ,Δ)*-M-matrix if* $\mathbb{A}^T$ *is a row-wise* (δ,Δ)*-M-matrix.*

It is well known that M-matrices are invertible matrices. The goal of Lemma 3.5 and of Corollary 3.6 is to get uniform estimates on the inverse of a uniform M-matrix. Let us stress that the estimates we obtain are far from being optimal in the applications we have in mind, namely the Finite Volume discretization of Richards' equation.

Lemma 3.5. *Let* $\mathbb{A} \in \mathcal{M}_N(\mathbb{R})$ *be a row-wise* (δ,Δ)*-M-matrix, then there exists* C *depending only on* δ,Δ *and* N *such that* $\|\mathbb{A}^{-1}\|_\infty \le C$.

Proof. For the ease of reading, we denote $\|\cdot\|$ instead of $\|\cdot\|_\infty$, and $\mathcal{I}_\delta$ instead of $\mathcal{I}_\delta(\mathbb{A})$. The property $\|\mathbb{A}^{-1}\| = \frac{1}{\alpha}$ is equivalent to

$$\frac{1}{\|\mathbb{A}^{-1}\|} = \max_{\|\boldsymbol{z}\|=1} \|\mathbb{A}\boldsymbol{z}\| = \alpha. \tag{59}$$

Let $\boldsymbol{z} = (z_i)_{1\le i\le N} \in \mathbb{R}^N$ with $\|\boldsymbol{z}\| = 1$. We assume, without loss of generality that there exists $i \in \{1,\dots N\}$ such that $z_i = 1$. In the sequel, we denote by $\mathcal{L}$ the maximum length of a transmissive path, i.e.,

$$\mathcal{L} := \max_{j\in\mathcal{I}_\delta(\mathbb{A})^c} \min_{j\in\mathcal{I}_\delta(\mathbb{A})} \text{length}\left(\mathcal{P}(i,j)\right) \le N-1. \tag{60}$$

Assume first that $i \in \mathcal{I}_\delta$, then $(\mathbb{A}\boldsymbol{z})_i = \sum_j a_{ij}z_j \ge \delta$, hence $\|\mathbb{A}^{-1}\| \le \frac{1}{\delta}$. Assume now that $i \in \mathcal{I}_\delta^c$, and let $\{k_p,\ 0 \le p \le L\}$ be a transmissive path with $k_0 = i$, $k_L \in \mathcal{I}_\delta$ and $L \le \mathcal{L}$. Denote by

$$c_p = \left(\left(\frac{\Delta}{\delta}\right)^p - 1\right)\frac{1}{\Delta-\delta}.$$

Let us show by induction that

$$z_{k_p} \ge 1 - c_p\alpha \qquad \forall p \in \{0,\dots,L\}. \tag{61}$$

Since $z_{k_0} = 1$, the relation (61) holds for $p = 0$. Now suppose that (61) holds for some $p \in \{0,\dots,N-1\}$. One knows from (59) that $\sum_{j=1}^N a_{k_pj}z_j \le \alpha$, hence, using (61), that $\sum_{j=1}^N a_{k_pj} \ge 0$, and that $a_{k_pj}z_j \ge a_{k_pj}$ for all $j \notin \{p,p+1\}$, one gets that

$$-a_{k_pk_{p+1}}(1 - z_{k_{p+1}}) \le \alpha(1 + c_p a_{k_pk_p}).$$

Since $a_{k_pk_p} \le \Delta$ and $a_{k_pk_{p+1}} \le -\delta$, this leads to

$$z_{k_{p+1}} \ge 1 - \alpha\frac{1+\Delta c_p}{\delta} = 1 - \alpha c_{p+1},$$

so that the proof of (61) is complete. As a consequence, one gets that $z_{k_L} \ge 1-\alpha c_L$, and thus that

$$a_{k_Lk_L}(1 - c_L\alpha) + \sum_{j\ne k_L} a_{k_Lj} \le \alpha.$$

Using that $k_L \in \mathcal{I}_\delta$, we obtain that $(\Delta c_L + 1)\,\alpha \ge \delta$. Therefore,

$$\alpha \ge \frac{1}{c_{L+1}} \ge \frac{1}{c_{\mathcal{L}+1}}$$

since the length L of the path $\{k_0,\dots k_L\}$ is bounded by $\mathcal{L}$ defined by (60). □

Corollary 3.6. *Let* $\mathbb{A} \in \mathcal{M}_N(\mathbb{R})$ *be a column-wise* (δ,Δ)*-M-matrix, then there exists* C *depending only on* δ,Δ *and* N *such that* $\|\mathbb{A}^{-1}\|_1 \le C$.

3.2. **A uniform non-degeneracy result.**

Proposition 3.7. *Assume that s and u satisfy he non-degeneracy condition (11), then there exist C_2 depending only on $\alpha^\star, \mathbf{g}, \|\lambda'\|_\infty, \Delta t^n$ and $\mathcal{T}$, and C_3 depending only on $\alpha_\star, \alpha^\star, \Delta t^n$ and $\mathcal{T}$ such that*

$$\|\mathbb{J}_{\mathcal{F}_n}(\boldsymbol{\tau})\|_1 \leq C_2, \qquad \| \left[\mathbb{J}_{\mathcal{F}_n}(\boldsymbol{\tau})\right]^{-1} \|_1 \leq C_3, \qquad \forall \boldsymbol{\tau} \in \mathbb{R}^{\#\mathcal{T}}.$$

Proof. Let us first make the Jacobian matrix $\mathbb{J}_{\mathcal{F}_n}(\boldsymbol{\tau}) = (j^n_{KL}(\boldsymbol{\tau}))_{K,L\in\mathcal{T}}$ explicit:

$$j^n_{KK}(\boldsymbol{\tau}) = s'(\tau_K) + \frac{\Delta t^n}{m_K} \sum_{\sigma\in\mathcal{E}_K} \left(m_\sigma g^+_{K,\sigma} \lambda'(s(\tau_K)) s'(\tau_K) + A_\sigma u'(\tau_K) \right),$$

$$j^n_{LK}(\boldsymbol{\tau}) = -\frac{\Delta t^n}{m_K} \left(m_\sigma g^+_{K,\sigma} \lambda'(s(\tau_K)) s'(\tau_K) + A_\sigma u'(\tau_K) \right) \quad \text{where } \sigma = K|L.$$

Proving that $\|\mathbb{J}_{\mathcal{F}_n}\|_1 \leq C_2$ is easy since all the coordinates of $\mathbb{J}_{\mathcal{F}_n}$ are uniformly bounded w.r.t. $\boldsymbol{\tau}$ thanks to the upper bound (12) on s' and u'.

Let us now prove that $\| \left[\mathbb{J}_{\mathcal{F}_n}(\boldsymbol{\tau})\right]^{-1} \|_1 \leq C_3$. Thanks to Corollary 3.6, it suffices to check that $\mathbb{J}_{\mathcal{F}_n}$ is a column-wise (δ, Δ)-M-matrix where δ and Δ depend on the prescribed quantities. First, it is easy to check that

$$j^n_{LK}(\boldsymbol{\tau}) \leq 0 \text{ if } L \neq K, \quad \sum_{L\in\mathcal{T}} j^n_{LK}(\boldsymbol{\tau}) \geq 0, \quad \text{and} \quad 0 \leq \delta_1 \leq j^n_{KK}(\boldsymbol{\tau}) \leq \Delta, \tag{62}$$

where

$$\delta = \alpha_\star \min\left\{1; \Delta t^n \min_K \left(\frac{\min_{\sigma\in\mathcal{E}_K} A_\sigma}{m_K} \right) \right\},$$

$$\Delta = \alpha^\star \max_{K\in\mathcal{T}} \left(1 + \frac{\Delta t^n}{m_K} \sum_{\sigma\in\mathcal{E}_K} (m_\sigma g^+_{K,\sigma} \|\lambda'\|_\infty + A_\sigma) \right).$$

Therefore, Condition (i) in Definition 3.4 is fulfilled.

It follows from the non-degeneracy condition (11) that for all $\boldsymbol{\tau} \in \mathbb{X}_{\mathcal{T},\text{int}}$ and all $K \in \mathcal{T}$, either $s'(\tau_K) \geq \alpha_\star$ or $u'(\tau_K) \geq \alpha_\star$. Let K be such that $s'(\tau_K) \geq \alpha_\star$, then

$$\sum_{L\in\mathcal{T}} j^n_{LK}(\boldsymbol{\tau}) \geq \alpha_\star \geq \delta,$$

whence $K \in \mathcal{I}_\delta(\mathbb{J}_{\mathcal{F}_n}(\boldsymbol{\tau})^T)$. On the other hand, if $u'(\tau_K) \geq \alpha_\star$ and $\mathcal{E}_{K,\text{ext}} \neq \emptyset$, then

$$\sum_{L\in\mathcal{T}} j^n_{LK}(\boldsymbol{\tau}) \geq \alpha_\star \frac{\Delta t^n}{m_K} \sum_{\sigma\in\mathcal{E}_{K,\text{ext}}} A_\sigma \geq \delta.$$

As a consequence, if $K \notin \mathcal{I}_\delta(\mathbb{J}_{\mathcal{F}_n}(\boldsymbol{\tau})^T)$, one has necessarily that $u'(\tau_K) \geq \alpha_\star$ and $\mathcal{E}_{K,\text{ext}} = \emptyset$. But in this case,

$$j^n_{LK}(\boldsymbol{\tau}) \leq -\alpha_\star \frac{\Delta t^n A_\sigma}{m_K} \leq -\delta \quad \text{if } \sigma = K|L \in \mathcal{E}_{K,\text{int}}.$$

The matrix $\mathbb{D}^n \in \mathbb{R}^{\#\mathcal{T}\times\#\mathcal{T}}$ defined by

$$\mathbb{D}^n_{KK} = \alpha_\star \frac{\Delta t^n}{m_K} \sum_{\sigma\in\mathcal{E}_K} A_\sigma, \qquad \mathbb{D}^n_{KL} = \begin{cases} -\alpha_\star \frac{\Delta t^n}{m_K} A_\sigma & \text{if } \sigma = K|L \\ 0 & \text{otherwise} \end{cases}$$

is irreducible and admits δ-transmissive paths from K to L for all $(K, L) \in \mathcal{T}^2$. This ensures that $\mathbb{J}_{\mathcal{F}_n}(\boldsymbol{\tau})$ admits a δ-transmissive path from any cell $K \in \mathcal{I}_\delta(\mathbb{J}_{\mathcal{F}_n}(\boldsymbol{\tau})^T)^c$

to any cell $L \in \mathcal{I}_\delta(\mathbb{J}_{\mathcal{F}_n}(\boldsymbol{\tau})^T)$. Therefore, Condition (ii) of Definition 3.4 is fulfilled. $\mathbb{J}_{\mathcal{F}_n}(\boldsymbol{\tau})^T$ is then a row-wise (δ,Δ)-M-matrix, thus $\mathbb{J}_{\mathcal{F}_n}(\boldsymbol{\tau})$ is a column-wise (δ,Δ)-M-matrix. □

The following corollary is a straightforward compilation of Newton-Kantorovich theorem 3.1 with Proposition 3.7.

Corollary 3.8 (local convergence of Newton's method)**.** *There exists $\rho^n > 0$ such that the Newton method (53) converges as soon as $\|\tau^{n,0} - \tau^n\| \leq \rho^n$.*

Remark 3.9 (global convergence for small enough time steps)**.** *The radius ρ^n appearing in Corollary 3.8 can be estimated thanks to [29] as soon as the second order derivatives s'' and u'' of s and u are uniformly bounded. It appears in a fairly natural way that ρ^n is a non-decreasing function of Δt^n. Then choosing Δt^n small enough, the variation between the time steps t^{n-1} and t^n is small and*

$$\|\boldsymbol{\tau}^{n-1} - \boldsymbol{\tau}^n\|_1 = \|\boldsymbol{\tau}^{n,0} - \boldsymbol{\tau}^n\|_1 \leq \rho^n.$$

Therefore, the convergence of the Newton method is ensured if one uses an adaptive time-step algorithm (as for instance in [10]).

3.3. **Control of the error induced by the inexact Newton procedure.** Assume that $s(\boldsymbol{\tau}^{n-1})$ is exactly known, then the exact solution $\boldsymbol{\tau}^n$ is obtained as the limit of $\left(\boldsymbol{\tau}^{n,k}\right)_k$. Computing the exact value of $\boldsymbol{\tau}^n$ is impossible and a convenient criterion must be adopted in order to stop the iterative procedure. This yields errors that accumulate along time. The goal of this section is to quantify the error induced by the inexact resolution of the nonlinear system.

3.3.1. *One step error estimates.* In this section, we assume that $s(\boldsymbol{\tau}^{n-1})$ is exact, and we want to quantify the error corresponding to one single iteration. In what follows, we consider the following residual based stopping criterion:

$$\text{(63)}\quad \text{stop the iterative procedure (53) if } \left\|\mathcal{F}_n(\boldsymbol{\tau}^{n,k})\right\|_1 = \sum_{K\in\mathcal{T}} \left|f_K(\tau_K^{n,k})\right| \leq \epsilon \Delta t^n$$

for some prescribed tolerance $\epsilon > 0$. We denote by $\boldsymbol{\tau}_\epsilon^n = \boldsymbol{\tau}^{n,k}$ when (63) is fulfilled and the loop is stopped. Then the non-degeneracy (58) of $[\mathbb{J}_{\mathcal{F}_n}]^{-1}$ provides directly the following error estimate:

$$\|\boldsymbol{\tau}_\epsilon^n - \boldsymbol{\tau}^n\|_1 \leq C_3 \epsilon \Delta t^n.$$

More than in the variable τ, whose physical sense is unclear, we are interested in evaluating the error on the reconstructed saturation profile. Thanks to the Lipschitz continuity of s —recall the non-degeneracy assumption (11)—, we have

$$\text{(64)}\qquad \|\pi_{\mathcal{T}} s(\boldsymbol{\tau}_\epsilon^n) - \pi_{\mathcal{T}} s(\boldsymbol{\tau}^n)\|_{L^1(\Omega)} \leq C_4 \epsilon \Delta t^n,$$

where $C_4 = \alpha^\star(\max_K m_K) C_3$.

3.3.2. *Quantification of the error accumulation.* In the previous paragraph, it was assumed that $s(\boldsymbol{\tau}^{n-1})$ was exactly known. In practice, one may consider that we know $s(\boldsymbol{\tau}^0)$ exactly, but merely the approximation of $s(\boldsymbol{\tau}_\epsilon^n)$ obtained after stopping the Newton iterative procedure after a finite number of iterations. Denote by $(s(\boldsymbol{\tau}^n))_{1\leq n\leq N}$ the iterated exact solutions to the scheme (31)–(32), and by

$(s(\boldsymbol{\tau}_\epsilon^n))_{1\leq n\leq N}$ the iterated inexact solutions obtained *via* the Newton method with the stopping criterion (63). Define $\mathcal{F}_{n,\epsilon} : \mathbb{X}_{\mathcal{T},\mathrm{int}} \to \mathbb{X}_{\mathcal{T},\mathrm{int}}$ by

$$\mathcal{F}_{n,\epsilon}(\boldsymbol{\tau}) = \mathcal{F}_n(\boldsymbol{\tau}) + s(\boldsymbol{\tau}^{n-1}) - s(\boldsymbol{\tau}_\epsilon^{n-1}). \tag{65}$$

In particular, one has $\mathbb{J}_{\mathcal{F}_{n,\epsilon}}(\boldsymbol{\tau}) = \mathbb{J}_{\mathcal{F}_n}(\boldsymbol{\tau})$ for all $\boldsymbol{\tau}$ and the results of §3.2 still hold for $\mathcal{F}_{n,\epsilon}$ instead of $\mathcal{F}_n$. The inexact Newton method then writes

(1) `Initialization:` define $\boldsymbol{\tau}_\epsilon^0 = \boldsymbol{\tau}^0$ by (28).
(2) `From` t^{n-1} `to` t^n`:`
   (a) set $\boldsymbol{\tau}_\epsilon^n = \boldsymbol{\tau}_\epsilon^{n-1}$
   (b) iterate Newton's algorithm until $\|\mathcal{F}_{n,\epsilon}(\boldsymbol{\tau}_\epsilon^n)\|_1 \leq \epsilon\Delta t^n$

Then, as claimed by the following statement, the $L^1(\Omega)$ error on the reconstructed saturation growth at most linearly with time. In particular, no exponential amplification of the error occurs in this context.

Proposition 3.10. *Let $(\boldsymbol{\tau}^n)_n$ be the exact solution to the scheme (31)–(32), and let $(\boldsymbol{\tau}_\epsilon^n)_n$ be the approximate solution computed by the inexact Newton method, then*

$$\|\pi_{\mathcal{T}} s(\boldsymbol{\tau}_\epsilon^n) - \pi_{\mathcal{T}} s(\boldsymbol{\tau}^n)\|_{L^1(\Omega)} \leq C_4\epsilon t^n, \qquad \forall n \in \{0,\dots,N\}, \tag{66}$$

where C_4 is introduced was (64).

Proof. We perform the proof by induction. Estimate (66) clearly holds for $n = 0$ since the initialization is exact. Now assume that it holds for $n-1$. Denote by $\widetilde{\boldsymbol{\tau}}_\epsilon^n$ the exact solution of the system $\mathcal{F}_{n,\epsilon}(\widetilde{\boldsymbol{\tau}}_\epsilon^n) = \mathbf{0}$. Then it follows from the discussion carried out in §3.3.1 that

$$\|\pi_{\mathcal{T}} s(\boldsymbol{\tau}_\epsilon^n) - \pi_{\mathcal{T}} s(\widetilde{\boldsymbol{\tau}}_\epsilon^n)\|_{L^1(\Omega)} \leq C_4\epsilon\Delta t^n.$$

On the other hand, applying Lemma 2.5, one gets that

$$\|\pi_{\mathcal{T}} s(\widetilde{\boldsymbol{\tau}}_\epsilon^n) - \pi_{\mathcal{T}} s(\boldsymbol{\tau}^n)\|_{L^1(\Omega)} \leq \|\pi_{\mathcal{T}} s(\boldsymbol{\tau}_\epsilon^{n-1}) - \pi_{\mathcal{T}} s(\boldsymbol{\tau}^{n-1})\|_{L^1(\Omega)} \leq C_4\epsilon t^{n-1}.$$

One concludes thanks to the triangle inequality. □

4. Numerical validation of the approach

In the previous section, we have shown that as long as the parametrization $\tau \mapsto (s(\tau), u(\tau))$ satisfies the condition (11), the Newton's method applied to the problem (52) exhibits local convergence. Moreover, due to Proposition 3.10, the error resulting from an inexact Newton's method can be efficiently controlled. Remark that the constant C_4 in (66) depends on the nonlinearities S and η only through the quantities $\alpha_\star$ and $\alpha^\star$. Therefore, the estimate (66) is robust with respect to the hydrodynamic properties of the soil. We illustrate this fact by numerical experiments presented below. As an example of mobility/capillary pressure relations we consider a popular Brooks-Corey model [6] for which we compare the efficiency of Newton's method resulting from the parametrization $u(\tau) = \tau$ and the one satisfying (11) with $\alpha_\star = \alpha^\star = 1$.

Let $p_b < 0$ and $\beta > 0$, in Brooks-Corey model the saturation and the mobility functions are given by

$$S(p) = \begin{cases} \left(\dfrac{p}{p_b}\right)^{-\beta} & \text{if } p < p_b, \\ 1 & \text{if } p \geq p_b, \end{cases}$$

and

$$\lambda(s) = s^{3+\frac{2}{\beta}},$$

providing in terms of Kirchhoff transform

$$\tilde{S}(u) = \begin{cases} \left(\dfrac{u}{u_b}\right)^{\frac{1}{\eta}} & \text{if } u < u_b, \\ 1 & \text{if } u \geq u_b, \end{cases}$$

with $\eta = \beta + 3 + \dfrac{1}{\beta}$ and $u_b = -\dfrac{p_b}{\beta\eta}$.

Remark that the parametrization based on $u(\tau) = \tau$ and $s(\tau) = \tilde{S}(u(\tau))$, referred as $u-$formulation, do not satisfy (11) since the derivative of $\tilde{S}(u)$ is singular at $u = 0$. As an alternative we consider the parametrization $\tau \mapsto (s(\tau), u(\tau))$ defined by the equation $\max(s'(\tau), u'(\tau)) = 1$ and the condition $s(0) = 0$, to which we refer as $\tau-$formulation. We obtain the following explicit formulas

$$s(\tau) = \begin{cases} \tau & \text{if } \tau < \tau_\star, \\ S(\tau - \tau_\star + u_b\tau_\star^\eta) & \text{if } \tau \geq \tau_\star, \end{cases}$$

and

$$u(\tau) = \begin{cases} u_b\tau^\eta & \text{if } \tau < \tau_\star, \\ \tau - \tau_\star + u_b\tau_\star^\eta & \text{if } \tau \geq \tau_\star, \end{cases}$$

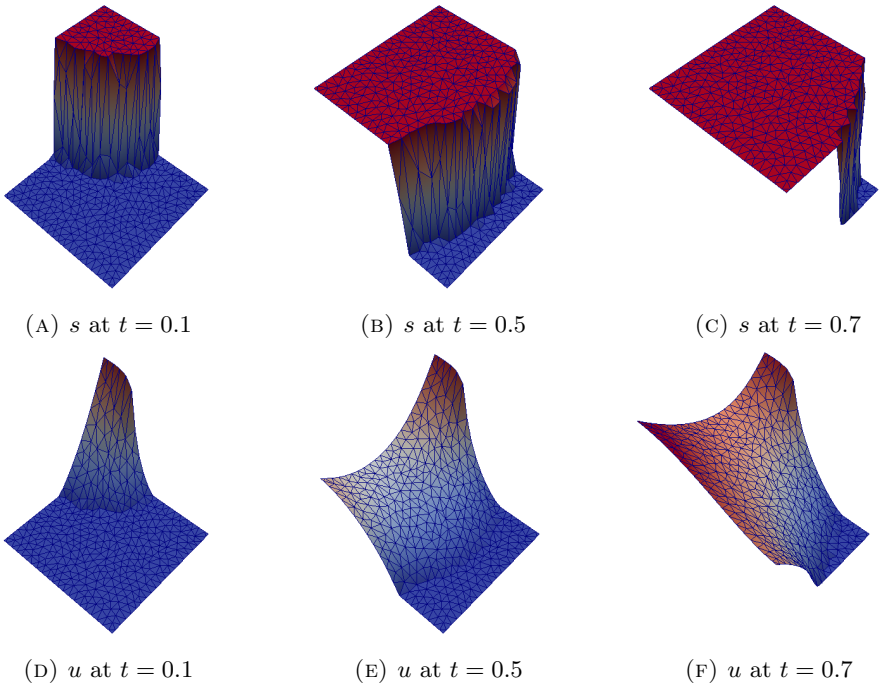
with $\tau_\star = \min\left((\eta u_b)^{\frac{1}{1-\eta}}, 1\right)$.

4.1. **First test case.** We consider a bidimensional porous domain $\Omega = (0,1) \times (0,1)$, which is initially very dry with $s_0(\boldsymbol{x}) = 10^{-6}$ in Ω. The water is injected at the pressure $p_D = 1$ through the portion of an upper boundary $\Gamma_D = \{(x_1, x_2) \mid x_1 \in (0, 0.3), x_2 = 1\}$. The gravity vector is given by $\mathbf{g} = -\nabla x_2$ and a zero flux boundary condition is prescribed on $\partial\Omega \setminus \Gamma_D$. The computations are performed on two quasi-uniform space discretizations of Ω composed of 396 and 1521 Voronoï cells referred as Mesh 1 and Mesh 2. The final time is set to $T = 0.7$ and the time step is equal to 0.01. Figure 2 shows, for $\beta = 4$ and $p_b = -10^{-2}$, the distribution of saturation and generalized pressure u at different times. The results are visualized on triangular Delaunay mesh dual to Mesh 1.

In order to challenge the robustness of both formulations, we set $p_b = -10^{-2}$ and we let the parameter β take values in the set $\{1, 2, 4, 8, 16\}$. For each value of β we compute, using $\tau-$formulation and tolerance $\epsilon_{ref} = 10^{-12}$, the reference solution denoted by $\left(\boldsymbol{\tau}_\beta^n\right)_{n\in\{1,\dots,N\}} \in \mathbb{X}_\mathcal{D}$. Then, for both formulations and for the values of $\epsilon \in \{10^{-2}, 10^{-4}, 10^{-6}, 10^{-8}, 10^{-10}, 10^{-12}\}$, we perform the calculations measuring the total number of Newton's iteration and the deviation, in the discrete $L^\infty(L^1)$ norm, of the "observable" variables u and s from the reference solution.

For a given value of β and of the tolerance ϵ, we denote by $\left(\overline{\boldsymbol{\tau}}_{\beta,\epsilon}^n\right)_{n\in\{1,\dots,N\}} \in \mathbb{X}_\mathcal{D}$ and $\left(\boldsymbol{\tau}_{\beta,\epsilon}^n\right)_{n\in\{1,\dots,N\}} \in X_\mathcal{D}$ the approximate solution of (52) obtained using the $u-$formulation and $\tau-$formulation respectively. The error produced by inexact Newton's method is measured by the quantities

$$\overline{err}_{\beta,\epsilon}^u = \frac{\|\pi_\mathcal{D}\overline{\boldsymbol{\tau}}_{\beta,\epsilon}^n - \pi_\mathcal{D}u(\boldsymbol{\tau}_\beta^n)\|_{L^\infty(0,T;L^1(\Omega))}}{\|\pi_\mathcal{D}u(\boldsymbol{\tau}_\beta^n)\|_{L^\infty(0,T;L^1(\Omega))}}$$



(A) s at $t = 0.1$ (B) s at $t = 0.5$ (C) s at $t = 0.7$

(D) u at $t = 0.1$ (E) u at $t = 0.5$ (F) u at $t = 0.7$

FIGURE 2. Snapshots of the reference solution at different times.

and

$$\overline{err}^{s}_{\beta,\epsilon} = \frac{\|\pi_{\mathcal{D}}\tilde{S}(\overline{\boldsymbol{\tau}}^{n}_{\beta,\epsilon}) - \pi_{\mathcal{D}}s(\boldsymbol{\tau}^{n}_{\beta})\|_{L^{\infty}(0,T;L^{1}(\Omega))}}{\|\pi_{\mathcal{D}}s(\boldsymbol{\tau}^{n}_{\beta})\|_{L^{\infty}(0,T;L^{1}(\Omega))}}$$

for $u-$formulation, and

$$err^{u}_{\beta,\epsilon} = \frac{\|\pi_{\mathcal{D}}u(\boldsymbol{\tau}^{n}_{\beta,\epsilon}) - \pi_{\mathcal{D}}u(\boldsymbol{\tau}^{n}_{\beta})\|_{L^{\infty}(0,T;L^{1}(\Omega))}}{\|\pi_{\mathcal{D}}u(\boldsymbol{\tau}^{n}_{\beta})\|_{L^{\infty}(0,T;L^{1}(\Omega))}}$$

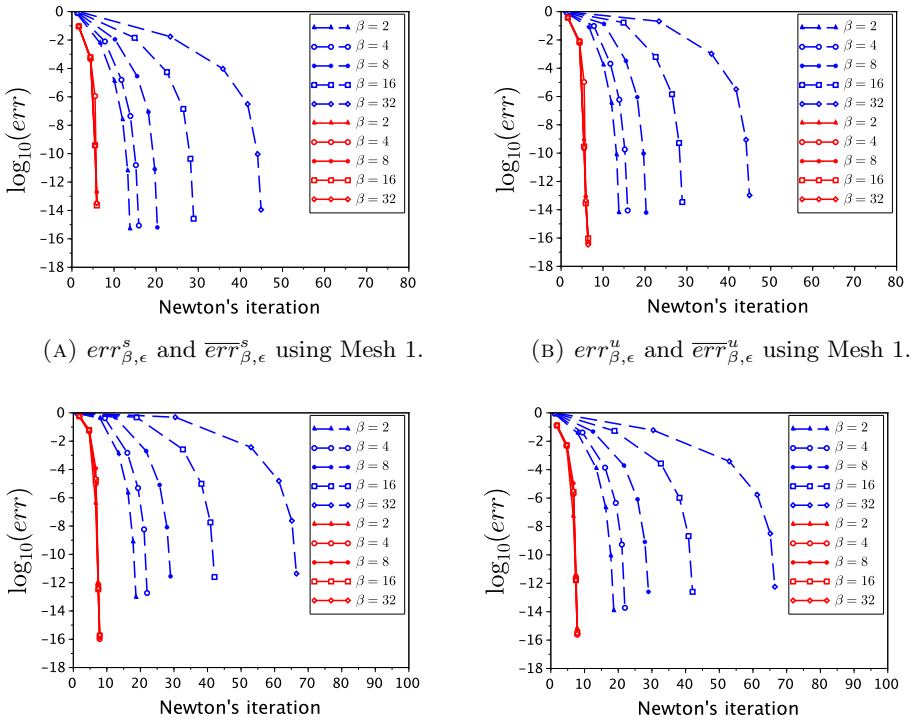
and

$$err^{s}_{\beta,\epsilon} = \frac{\|\pi_{\mathcal{D}}s(\boldsymbol{\tau}^{n}_{\beta,\epsilon}) - \pi_{\mathcal{D}}s(\boldsymbol{\tau}^{n}_{\beta})\|_{L^{\infty}(0,T;L^{1}(\Omega))}}{\|\pi_{\mathcal{D}}s(\boldsymbol{\tau}^{n}_{\beta})\|_{L^{\infty}(0,T;L^{1}(\Omega))}}$$

for $\tau-$formulation.

Figure 3 exhibits, for Meshes 1 and 2, the behavior of the relative error $err^{\xi}_{\beta,\epsilon}$ and $\overline{err}^{\xi}_{\beta,\epsilon}$, $\xi = u, s$ as the function of an average number of iterations per time step required by Newton's method in order to converge up to the given tolerance ϵ. We observe that in order to achieve the same precision $u-$formulation require a much larger number of iterations then $\tau-$formulation. Moreover the number of Newton's iterations, for $u-$formulation, increases with β, whereas τ-formulation remains robust with respect to this parameter. The contrast in the efficiency of two formulations is amplified as the mesh is refined.

4.2. **Second test case.** The goal of this test case is to give a numerical evidence that the inexact Newton's method applied to the u-formulation produces large



(A) $err^s_{\beta,\epsilon}$ and $\overline{err}^s_{\beta,\epsilon}$ using Mesh 1.

(B) $err^u_{\beta,\epsilon}$ and $\overline{err}^u_{\beta,\epsilon}$ using Mesh 1.

(C) $err^s_{\beta,\epsilon}$ and $\overline{err}^s_{\beta,\epsilon}$ using Mesh 2.

(D) $err^u_{\beta,\epsilon}$ and $\overline{err}^u_{\beta,\epsilon}$ using Mesh 2.

FIGURE 3. Relative error as the function of the average number of Newton's iterations per time step using $\tau-$formulation (red solid lines) and $u-$formulation (blue dashed lines).

errors due to troubles in the conservation of mass. To do so, we prescribe a zero-flux condition on the whole boundary of $\Omega = (0,1) \times (0,1)$, whereas the initial saturation satisfies

$$s_0 = \begin{cases} 0.5 & \text{in} \quad \Omega', \\ 10^{-6} & \text{in} \quad \Omega \setminus \Omega' \end{cases}$$

with $\Omega' = \{(x_1, x_2) \mid x_1 < 0.5 \text{ and } x_2 > 0.5\}$. We set $\beta = 4$ and $p_b = -10^{-2}$. The effects of gravity neglected so that the flow is only driven by diffusion. Figure 5 exhibits, for different times, the saturation field associated with the reference solution, which is computed using $\tau-$formulation and the tolerance $\epsilon_{ref} = 10^{-16}$. Since the flow is very slow the large time steps are needed, we set $T = 10^5$ and $\Delta t = 10^3$. Computations are performed using Mesh 1. Let $M = \int_\Omega s_0 \mathrm{d}\boldsymbol{x}$, we define the relative mass conservation error by

$$\overline{mass\ err}^s_{\beta,\epsilon} = \frac{1}{M} \max_{n \in \{1,\dots,N\}} \left| \int_\Omega \pi_{\mathcal{D}} \tilde{S}(\overline{\boldsymbol{\tau}}^n_{\beta,\epsilon}) \mathrm{d}\boldsymbol{x} - M \right|$$

and

$$mass\ err^s_{\beta,\epsilon} = \frac{1}{M} \max_{n \in \{1,\dots,N\}} \left| \int_\Omega \pi_{\mathcal{D}} s(\boldsymbol{\tau}^n_{\beta,\epsilon}) \mathrm{d}\boldsymbol{x} - M \right|.$$

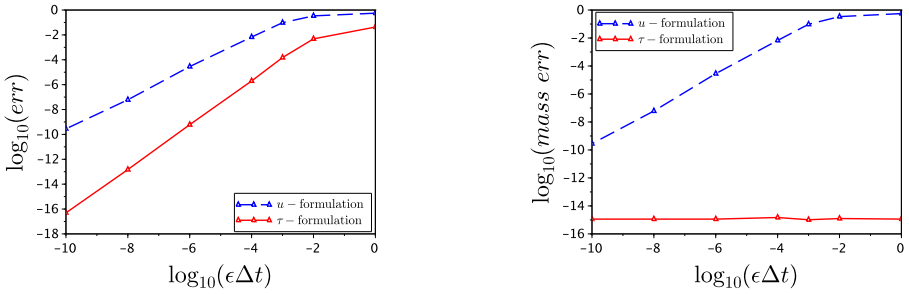


FIGURE 4. At left $err^s_{\beta,\epsilon}$ (red solid lines) and $\overline{err}^s_{\beta,\epsilon}$ (blue dashed lines) as the function of ϵ, at right the relative mass conservation error as the function of ϵ.

Figure 4 exhibits the $L^\infty(L^1)$ relative saturation error $err^s_{\beta,\epsilon}$, $\overline{err}^s_{\beta,\epsilon}$, and the relative mass conservation error $mass\ err^s_{\beta,\epsilon}$ and $\overline{mass\ err}^s_{\beta,\epsilon}$ as the functions of ϵ. As one can see the $L^\infty(L^1)$ error produced by $u-$formulation is dominated by the mass conservation error. Remark that even for the rather small values $\epsilon = 10^{-6}$ or $\epsilon = 10^{-7}$, the error produced by $u-$formulation is still significant (see Figures 6). In contrast $\tau-$formulation leads to mush smaller errors and, for any value of ϵ, conservatives the mass up to a precision of order 10^{-15}.

The very high accuracy for mass conservation observed with the $\tau-$formulation can be explained as follows. With the values of the parameters β, p_b we have chosen, one has $s(\tau) = \tau$ for $\tau \in [0,1]$ and hence for all $\tau \in \Omega \times (0,T)$ in view of initial and boundary conditions. In addition, in view of (32) we have

$$\sum_{K\in\mathcal{T}} \sum_{\sigma\in\mathcal{E}_K} F_{K,\sigma}(\boldsymbol{\tau}^n) = 0, \qquad \forall \boldsymbol{\tau}^n \in \mathbb{X}_{\mathcal{D}}.$$

Therefore, at each step of inexact Newton's method, the flux contribution globally offset, hence we have

$$\sum_{K\in\mathcal{T}} m_K \left(s(\tau^{n,k}_{\epsilon,K}) + s'(\tau^{n,k}_{\epsilon,K})(\tau^{n,k+1}_{\epsilon,K} - \tau^{n,k}_{\epsilon,K}) - s(\tau^{n-1}_{\epsilon,K}) \right) = 0.$$

Since $\tau \mapsto s(\tau)$ is linear, one has $s(\tau^{n,k+1}_{\epsilon,K}) = s(\tau^{n,k}_{\epsilon,K}) + s'(\tau^{n,k}_K)(\tau^{n,k+1}_{\epsilon,K} - \tau^{n,k}_{\epsilon,K})$, which implies that the mass is exactly conserved (assuming that linear algebraic computations are exact) at each iteration of Newton's method.

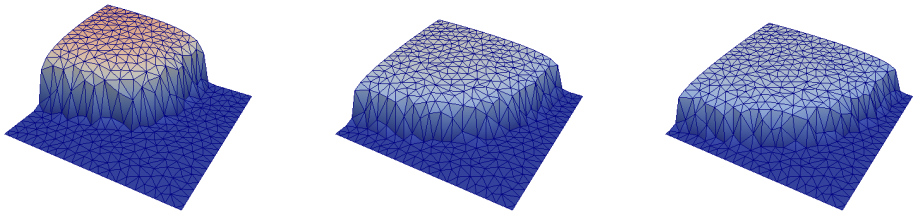
REFERENCES

[1] H. W. Alt and S. Luckhaus. Quasilinear elliptic-parabolic differential equations. *Math. Z.*, 183(3):311–341, 1983.

[2] B. Andreianov, C. Cancès, and A. Moussa. A nonlinear time compactness result and applications to discretization of degenerate parabolic-elliptic PDEs. HAL: hal-01142499, 2015.

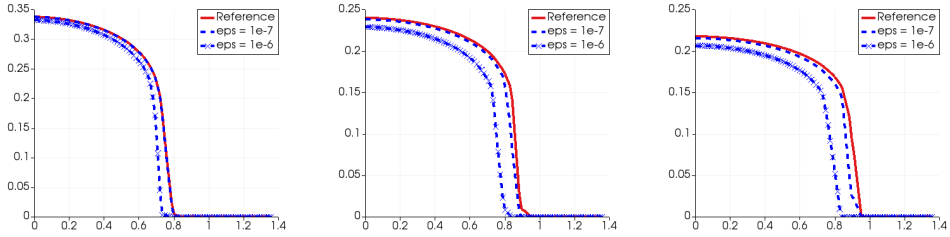
[3] J. Bear and Y. Bachmat. *Introduction to modeling of transport phenomena in porous media.* Kluwer Academic Publishers, Dordrecht, The Netherlands, 1990.

[4] L. Bergamaschi and M. Putti. Mixed finite elements and Newton-type linearizations for the solution of Richards' equation. *Int. J. Numer. Meth. Eng.*, 45(8):1025–1046, 1999.



(A) $t = 5\ 10^3$ (B) $t = 50\ 10^3$ (C) $t = 100\ 10^3$

FIGURE 5. Saturation field of the reference solution at different times.



(A) $t = 5\ 10^3$ (B) $t = 50\ 10^3$ (C) $t = 100\ 10^3$

FIGURE 6. Saturation profile of the reference solution along the line $y = 1 - x$ compared to the saturation profiles of the approximate solutions computed using u-formulation and $\epsilon = 10^{-8}$ and 10^{-9}.

[5] K. Brenner, M. Groza, L. Jeannin, R. Masson, and J. Pellerin. Immiscible two-phase darcy flow model accounting for vanishing and discontinuous capillary pressures: application to the flow in fractured porous media. *Proceedings of ECMOR XV*, 2016. To appear.

[6] R. H. Brooks and A. T. Corey. Hydraulic properties of porous media and their relation to drainage design. *Transactions of the ASAE*, 7(1):0026–0028, 1964.

[7] X.-C. Cai and D. E. Keyes. Nonlinearly preconditioned inexact Newton algorithms. *SIAM J. Sci. Comp.*, 24(1):183–200, 2002.

[8] C. Cancès and T. Gallouët. On the time continuity of entropy solutions. *J. Evol. Equ.*, 11(1):43–55, 2011.

[9] C. Cancès and C. Guichard. Convergence of a nonlinear entropy diminishing Control Volume Finite Element scheme for solving anisotropic degenerate parabolic equations. *Math. Comp.*, 85(298):549–580, 2016.

[10] C. Cancès and C. Guichard. Numerical analysis of a robust free energy-diminishing Finite Volume scheme for degenerate parabolic equations with gradient structure. preprint HAL: hal-01119735, accepted for publication in Found. Comput. Math., 2016.

[11] C. Cancès and M. Pierre. An existence result for multidimensional immiscible two-phase flows with discontinuous capillary pressure field. *SIAM J. Math. Anal.*, 44(2):966–992, 2012.

[12] J. Carrillo. On the uniqueness of the solution of the evolution dam problem. *Nonlinear Anal.*, 22(5):573–607, 1994.

[13] J. Carrillo. Entropy solutions for nonlinear degenerate problems. *Arch. Ration. Mech. Anal.*, 147(4):269–361, 1999.

[14] J. Carrillo. Conservation laws with discontinuous flux functions and boundary condition. *J. Evol. Equ.*, 3(2):283–301, 2003.

[15] F. H Clarke. *Optimization and nonsmooth analysis*. Siam, 1990.

[16] K. Deimling. *Nonlinear functional analysis*. Springer-Verlag, Berlin, 1985.

[17] H.-J.G. Diersch and P. Perrochet. On the primary variable switching technique for simulating unsaturated–saturated flows. *Adv. Water Resour.*, 23(3):271–301, 1999.
[18] V. Dolean, M. J. Gander, W. Kheriji, F. Kwok, and R. Masson. Nonlinear preconditioning: how to use a nonlinear Schwarz method to precondition Newton's method. preprint HAL:hal-01171167, July 2015.
[19] J. Droniou and R. Eymard. Uniform-in-time convergence of numerical methods for non-linear degenerate parabolic equations. *Numer. Math.*, 132(4):721–766, 2016.
[20] J. Droniou, R. Eymard, T. Gallouët, and R. Herbin. Gradient schemes: a generic framework for the discretisation of linear, nonlinear and nonlocal elliptic and parabolic equations. *Math. Models Methods Appl. Sci.*, 23(13):2395–2432, 2013.
[21] R. Eymard, T. Gallouët, M. Ghilani, and R. Herbin. Error estimates for the approximate solutions of a nonlinear hyperbolic equation given by finite volume schemes. *IMA J. Numer. Anal.*, 18(4):563–594, 1998.
[22] R. Eymard, T. Gallouët, C. Guichard, R. Herbin, and R. Masson. TP or not TP, that is the question. *Comput. Geosci.*, 18:285–296, 2014.
[23] R. Eymard, T. Gallouët, and R. Herbin. Finite volume methods. Ciarlet, P. G. (ed.) et al., in Handbook of numerical analysis. North-Holland, Amsterdam, pp. 713–1020, 2000.
[24] R. Eymard, T. Gallouët, and R. Herbin. Discretization of heterogeneous and anisotropic diffusion problems on general nonconforming meshes sushi: a scheme using stabilization and hybrid interfaces. *IMA J. Numer. Anal.*, 30(4):1009–1043, 2010.
[25] R. Eymard, M. Gutnic, and D. Hilhorst. The finite volume method for Richards equation. *Comput. Geosci.*, 3(3-4):259–294, 1999.
[26] J. Fuhrmann. Existence and uniqueness of solutions of certain systems of algebraic equations with off-diagonal nonlinearity. *Appl. Numer. Math.*, 37:359–370, 2001.
[27] J. Fuhrmann and H. Langmach. Stability and existence of solutions of time-implicit finite volume schemes for viscous nonlinear conservation laws. *Appl. Numer. Math.*, 37:201–230, 2001.
[28] G. Gagneux and M. Madaune-Tort. Unicité des solutions faibles d'équations de diffusion-convection. *C. R. Acad. Sci. Paris Sér. I Math.*, 318(10):919–924, 1994.
[29] W. B. Gragg and R. A. Tapia. Optimal error bounds for the Newton-Kantorovich theorem. *SIAM J. Numer. Anal.*, 11:10–13, 1974.
[30] W. Jäger and J. Kačur. Solution of porous medium type systems by linear approximation schemes. *Numer. Math.*, 60(3):407–427, 1991.
[31] W. Jäger and J. Kačur. Solution of doubly nonlinear and degenerate parabolic problems by relaxation schemes. *RAIRO Modél. Math. Anal. Numér.*, 29(5):605–627, 1995.
[32] L. V. Kantorovich. On Newtons method for functional equations. *Dokl. Akad. Nauk SSSR*, 59(7):1237–1240, 1948.
[33] F. Lehmann and P. H. Ackerer. Comparison of iterative methods for improved solutions of the fluid flow equation in partially saturated porous media. *Transport in Porous Media*, 31(3):275–292, 1998.
[34] J. Leray and J. Schauder. Topologie et équations fonctionnelles. *Ann. Sci. École Norm. Sup. (3)*, 51:45–78, 1934.
[35] F. List and F. A. Radu. A study on iterative methods for solving Richards equation. *Comput. Geosci.*, online first:1–13, 2016.
[36] S. Martin and J. Vovelle. Convergence of implicit finite volume methods for scalar conservation laws with discontinuous flux function. *ESAIM Math. Model. Numer. Anal.*, 42(5):699728, 2008.
[37] J. M. Ortega. The Newton-Kantorovich theorem. *Amer. Math. Monthly*, 75:658–660, 1968.
[38] J. M. Ortega and W. C. Rheinboldt. *Iterative solution of nonlinear equations in several variables*. Academic Press, New York-London, 1970.
[39] F. Otto. L^1-contraction and uniqueness for quasilinear elliptic-parabolic equations. *J. Differential Equations*, 131:20–38, 1996.
[40] I. S. Pop, F. Radu, and P. Knabner. Mixed finite elements for the Richards equation: linearization procedure. *J. Comput. Appl. Math.*, 168(1):365–373, 2004.
[41] L. Qi and J. Sun. A nonsmooth version of Newton's method. *Mathematical programming*, 58:353–367, 1993.

[42] F. A Radu, I. S. Pop, and P. Knabner. Newtontype methods for the mixed finite element discretization of some degenerate parabolic equations. In *Numerical mathematics and advanced applications*, pages 1192–1200. Springer, 2006.
[43] L. A. Richards. Capillary conduction of liquids through porous mediums. *Journal of Applied Physics*, 1(5):318–333, 1931.
[44] X. Wang and H. A. Tchelepi. Trust-region based solver for nonlinear transport in heterogeneous porous media. *J. Comput. Phys.*, 253:114–137, 2013.
[45] R. Younis, H. A. Tchelepi, and K. Aziz. Adaptively localized continuation-Newton method–nonlinear solvers that converge all the time. *SPE Journal*, 15(02):526–544, 2010.
[46] R. L. Zarba, E. T. Bouloutas, and M. Celia. General mass-conservative numerical solution for the unsaturated flow equation. *Water Resour. Res.*, 26(7):1483–1496, 1990.

Konstantin Brenner
Laboratoire Jean-Alexandre Dieudonné, Université de Nice Sophia Antipolis,
Team Coffee INRIA Sophia Antipolis Méditerranée,
06108 Nice Cedex 02, France.
konstantin.brenner@unice.fr

Clément Cancès
Team Rapsodi INRIA Lille - Nord Europe,
40, avenue Halley, 59650 Villeneuve d'Ascq, France.
clement.cances@inria.fr